(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
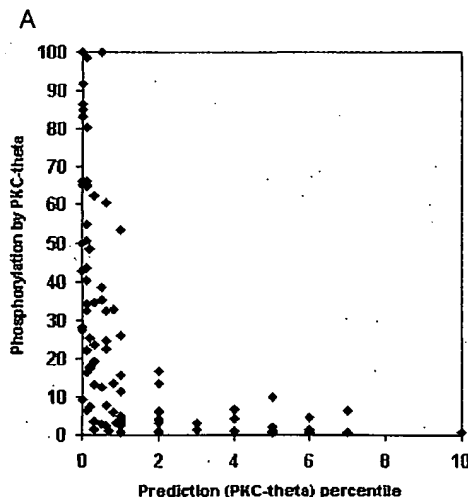31 March 2005 (31.03.2005)

PCT

(10) International Publication Number
WO 2005/028666 A2

(54) Title: DETERMINING KINASE SPECIFICITY

Prospective validation of scoring for PKC-theta PSSM

(57) Abstract: The invention provides methods, articles, software, kits as well as sets and arrays of peptides for determining the spectrum of peptidyl sequences that are recognized and phosphorylated by a kinase. The invention also provides binding entities that specifically distinguish phosphorylated and non-phosphorylated peptidyl sequences.

DETERMINING KINASE SPECIFICITY

This application claims priority from U.S. Application Ser. No.

5   10/660,370 filed September 11, 2003, the contents of which are incorporated

herein in their entireties.

Government Funding

The invention described herein was developed with support from the

National Institutes of Health. The U.S. Government has certain rights in the

10   invention.

Field of the Invention

The invention relates to methods, articles, software and kits for

determining the spectrum of peptidyl sequences that are recognized and

phosphorylated by a kinase, peptides that include kinase recognition sites and

15   binding entities that specifically distinguish phosphorylated versus non-

phosphorylated peptidyl sequences.

Background of the Invention

The activity of cells is regulated by external signals that stimulate or

inhibit intracellular events. The process by which stimulatory or inhibitory

20   signals are transmitted into and within a cell to elicit an intracellular response is

referred to as signal transduction. Proper signal transduction is essential for

proper cellular function. Defects in various components of signal transduction

pathways, from cell surface receptors to activators of gene transcription, account

for a vast number of diseases, including numerous forms of cancer, vascular

25   diseases and neuronal diseases.

Signal transduction is largely mediated by protein kinases. Protein

kinases are enzymes that phosphorylate other proteins and/or themselves (auto-

phosphorylation). A major rate-limiting problem in understanding signal

transduction within cells is to determine which kinase phosphorylates which

30   protein substrate at which sites within the protein substrate.

Eukaryotic protein kinases are numerous and diverse; there are more than

500 human genes than encode different protein kinases (Manning G et al. 2002.

Science 298:1912-1934). Eukaryotic protein kinases that are involved in signal transduction can be divided into three major groups based upon their substrate utilization. First, the protein-tyrosine specific kinases can phosphorylate substrates on tyrosine residues. Second, the protein-serine/threonine specific

5   kinases can phosphorylate substrates at serine and/or threonine residues. Finally, the dual-specificity kinases can phosphorylate substrates at tyrosine, serine and/or threonine residues.

In order to insure fidelity in intracellular signal transduction cascades it is essential that each protein kinase have exquisite specificity for its target

10   substrate(s). In general, kinases appear to phosphorylate multiple different target sites on multiple proteins, thereby allowing branching of an initial signal delivered to a cell in multiple directions in order to coordinate a set of events that occur in parallel for a given cellular response (see, for example, Roach, P. J. (1991) J. Biol. Chem. 266:14139-14142).

15   The substrate specificity of a protein kinase can be influenced by at least three general mechanisms that depend on the overall structure of the enzyme. First, specific domains in certain protein kinases can target the kinase to specific locations in the cell, thereby restricting the substrate availability of the kinase. Second, domains in the kinase, distinct from its catalytic domain, may provide

20   high affinity association with either the substrate or an adapter molecule that presents the substrate to the kinase. Finally, kinase specificity is ultimately provided by the structure of the catalytic site of the protein kinase that drives it to select one peptide substrate sequence over another.

Although the number of protein kinases that have been implicated in

25   intracellular signaling is quite large, detailed information about the sequence specificity of these kinases is available for only a limited number of these kinases. Shortcomings in the available approaches for detailed characterization of kinase specificity are largely responsible for this scarcity of information. One systematic approach to characterization of kinase specificity involves

30   collecting information on many specific substrates for a kinase and determining common features amongst the substrates sequences (Kreegipuu A et al. 1998. FEBS Lett 430:45-50). Such determination of the individual substrates is a laborious and largely empirical process, making this a slow and relatively inefficient way to derive comprehensive information on kinase specificity.

Serine/threonine kinases can be subdivided by peptide specificity into three broad classes: basophilic kinases that phosphorylate sites with clusters of positively charged amino acid residues, acidophilic kinases that phosphorylate sites with clusters of negatively charged amino acid residues and proline-

5   directed kinases that phosphorylate sites in which Ser/Thr is followed immediately by a proline (i.e. proline is at the P+1 position).

In the early 1990s, Cantley and colleagues invented a method that attempts to accurately predict the spectrum of good peptide substrates for a kinase (see U.S. Patent No. 5,532,167; Songyang et al. (1994) Curr. Biol. 4:973-

10   982). Predictions of substrate specificity made by this method are available at a website at scansite.mit.edu/. See also, Obenauer et al. (2003) Nucleic Acids Res. 31:3635-3641; Yaffe et al. (2001) Nat. Biotechnol. 19:348-353. Other workers have tested the specificities of kinases using one or more known substrates. See, Himpel et al. (2000) J. Biol. Chem. 275:2431-2438, Velentza et al. (2001) J.

15   Biol. Chem. 276:38956-38965; Dostmann et al. (1999) Pharmacol. Ther. 82:373-387; Tegge et al. (1998) Methods Mol Biol 87:99-106; Tegge et al. (1995) Biochemistry 34:10569-10577.

Limitations typical of these previous approaches include a failure to validate the substrate specificities indicated by the methods employed, a

20   propensity for seeking optimal substrate sequences rather than defining the universe of preferred substrates, and/or assumptions that a method provides general information when it may provide rather narrow information. Thus, there is a need for an alternative method to accurately characterize the universe of preferred substrates for kinases.

25

Summary of the Invention

The invention relates to determination of the range of substrate specificities of protein kinases, to prediction of sites on sequenced proteins that are most likely to be phosphorylated by each kinase studied, to visual

30   representation of those kinase specificities, to validation *in vitro* that peptides corresponding to those predicted sites are indeed phosphorylated by each kinase studied, and to validation of phosphorylation of those sites *in vivo*. The invention provides a simple and efficient method for determining the amino acid residue preferences for peptidyl sequences phosphorylated by a kinase, as well

as for predicting which sites will be preferentially phosphorylated by the kinase, and software that facilitates those methods. The invention also provides an informative graphical format for visually representing that information and software to output data in that format. Peptide sequences proven to be well

5   phosphorylated by protein kinase C are also provided.

In one embodiment, the invention provides a test set of peptide pools for identifying kinase substrate specificities. Such a test set for characterizing substrate specificities of kinases has at least two peptide pools. In general, substantially every peptide in each of the peptide pools includes one defined

10   phosphorylatable amino acid position, one query amino acid position, at least one anchor amino acid position, and at least one degenerate amino acid position. Substantially every peptide of every peptide pool has an identical phosphorylatable amino acid that can be phosphorylated by a kinase at the phosphorylatable amino acid position. The query amino acid position is at a

15   defined position relative to the phosphorylatable amino acid position within substantially every peptide of every peptide pool, but a query amino acid's identity at the query amino acid position is systematically varied from one peptide pool to the next peptide pool within the test set of peptide pools. Each anchor amino acid position is at a defined position relative to the

20   phosphorylatable amino acid position within substantially every peptide of every peptide pool and each anchor amino acid position has an identical anchor amino acid at that anchor amino acid position within every peptide of every peptide pool. Each degenerate amino acid position within every peptide of every peptide pool is occupied by an amino acid from a defined mixture of amino acids. In

25   some embodiments, the query amino acid position is not adjacent to an anchor amino acid position or the query amino acid position is not adjacent to the phosphorylatable amino acid position in any peptide pool of the test set. In some test sets of the invention, no anchor amino acid positions (or anchor amino acids) are present. However, such test sets do have a phosphorylatable amino acid

30   position, and at least one query amino acid position. Such "anchor-free" test sets will also generally have at least one degenerate amino acid position.

In other embodiments, the invention provides a test set like those described above except that every peptide of every peptide pool has an identical query amino acid but the position of the query amino acid relative to the

phosphorylatable amino acid position is systematically varied from one peptide pool to the next peptide pool within the test set of peptide pools. One desirable query amino acid to use in such a test set is arginine.

Another aspect of the invention is a test set for characterizing substrate
5   specificities of kinases that includes at least two peptide pools, wherein substantially every peptide in each of the peptide pools includes one phosphorylatable amino acid position, one query amino acid position, and at least one degenerate amino acid position, and wherein: (a) each peptide of every peptide pool has an identical phosphorylatable amino acid that can be
10   phosphorylated by a kinase at the phosphorylatable amino acid position; (b) the query amino acid position is at a defined position relative to the phosphorylatable amino acid position within every peptide of every peptide pool but a query amino acid's identity at the query amino acid position is systematically varied from one peptide pool to the next peptide pool within the
15   test set of peptide pools; (c) each degenerate amino acid position within every peptide of every peptide pool is occupied by an amino acid selected from a defined mixture of amino acids; and (d) the query amino acid position is not adjacent to the phosphorylatable amino acid position in any peptide pool of the test set. At least one degenerate position in each peptide pool in the test set can
20   be occupied by a defined mixture of more than five amino acids. Such a defined mixture can include all natural amino acids except cysteine. Alternatively, each amino acid's relative abundance in the defined mixture can be approximately that amino acid's relative abundance in the human proteome. In some embodiments, the defined mixture of amino acids includes arginine. Some of
25   the test sets of the invention have at least four peptide pools and each of the four peptide pools has a different query amino acid. Some of the test sets of the invention have a query amino acid position that is two positions N-terminal to the phosphorylatable amino acid position. Other test sets of the invention have a query amino acid position that is two positions C-terminal to the
30   phosphorylatable amino acid position. In some embodiments, one query amino acid of the test set is arginine. The peptide pool of the test sets of the invention can be a soluble mixture of peptides. Alternatively, substantially every peptide in each peptide pool is attached to a solid support. In some embodiments, substantially every peptide in each peptide pool is linked to biotin.

In other embodiments, the test sets of the invention are like those
described in the preceding paragraph but those test sets also have at least one
anchor amino acid position, wherein: (a) each anchor amino acid position is at a
defined position relative to the phosphorylatable amino acid position within

5   every peptide of every peptide pool and each anchor amino acid position has an
identical anchor amino acid at that anchor amino acid position within every
peptide of every peptide pool; and (b) the query amino acid position is not
adjacent to an anchor amino acid position in any peptide pool of the test set. In
some embodiments, at least one anchor amino acid is arginine. The anchor

10   amino acid position can be located one position C-terminal or one position N-
terminal to the phosphorylatable amino acid position. In other embodiments,
arginine is the anchor amino acid and the (arginine) anchor amino acid position
is located three positions N-terminal to the phosphorylatable amino acid
position. In some embodiments, every peptide in each of the peptide pools has

15   less than four anchor amino acids

Another aspect of the invention is a test set for characterizing substrate
specificities of kinases having at least two peptide pools, wherein every peptide
in each of the peptide pools comprises one phosphorylatable amino acid
position, one query amino acid, and at least one degenerate amino acid position,

20   and wherein: (a) each peptide of every peptide pool has an identical
phosphorylatable amino acid that can be phosphorylated by a kinase at the
phosphorylatable amino acid position; (b) every peptide of every peptide pool
has an identical query amino acid but the position of the query amino acid
relative to the phosphorylatable amino acid position is systematically varied

25   from one peptide pool to the next peptide pool within the test set of peptide
pools; and (c) each degenerate amino acid position within every peptide of every
peptide pool is occupied by an amino acid from a defined mixture of amino
acids. The query amino acid of this test set can be arginine. In this test set, each
peptide of every peptide pool can have at least one anchor amino acid position

30   that is at a defined position relative to the phosphorylatable amino acid position,
and each anchor amino acid position of peptides within a peptide pool can have
an identical anchor amino acid at that anchor amino acid position. In some
embodiments, the anchor amino acid of this test set is arginine and the anchor

amino acid position is two positions N-terminal to the phosphorylatable amino
acid position.

      Another aspect of the invention is a test set of peptides for characterizing
kinase substrate specificity that includes at least 50 separate peptides, each

5   peptide having a sequence of between 6 and 30 amino acids, wherein each
peptide sequence is different from every other peptide sequence, and wherein at
least 50 peptides have two or more arginines within 6 amino acid positions of a
serine or threonine. Such a test set can have at least 96 separate peptides that
each include two or more arginines within 6 amino acid positions of a serine or

10   threonine. In another embodiment, at least half of the peptides in the test set
have two or more arginines within 6 residues of a serine or threonine. In a
further embodiment, at least 50 peptides have two or more arginines but two of
these arginines are not within 2 to 3 positions N-terminal to the serine or
threonine. In some of the test sets of the invention, at least 50 peptides have

15   three or more arginine residues within 6 residues of a serine or threonine. One
or more lysine residues can also be included within 6 residues of a serine or
threonine in the peptides of the test set. Substantially every peptide in some of
the test sets of the invention corresponds to a peptidyl sequence in a mammalian
protein and the peptidyl sequence is within 30 amino acids of the protein's N-

20   terminus or C-terminus

      Another aspect of the invention is a peptide set comprising two or more
pools of peptides, wherein each pool has peptides having substantially identical
peptide sequences and the peptide sequences in each pool are selected from the
group consisting essentially of SEQ ID NO: 76, 81, 82, 87, 89-92, 94, 97-99,

25   102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143,
144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211,
213-216, 474-516 or 517.

      Another aspect of the invention is an isolated peptide having any one of
SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113,

30   121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171,
173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-516 or 517. A serine
or threonine in the peptide can be phosphorylated.

Another aspect of the invention is an isolated phosphorylated peptide having any one of SEQ ID NO: 298, 301-324,326-347, 349-400, 402-410, 412-473, 571-643 or 644.

Another aspect of the invention is an binding entity whose binding

5   differentiates between a peptide having any one of SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517 or 570, and the peptide after phosphorylation by protein kinase C theta; wherein the binding entity has

10   substantially no binding to a phosphorylated peptide having SEQ ID NO: 229 (WKN-pS-IRH). Many of the antibodies of the invention recognize phosphorylation sites at the N-termini and C-termini of mammalian proteins. In some embodiments, the binding entity binds with greater affinity to the peptide after phosphorylation than before phosphorylation. In other embodiments, the

15   binding entity binds with greater affinity to the peptide before phosphorylation than after phosphorylation. The binding entity can, for example, be an antibody, an antibody fragment or a mixture thereof. The peptide recognized by the binding entity can be part of a mammalian protein. In some embodiments, the peptide's sequence is within 30 amino acids of the protein's N-terminus or C-

20   terminus of said protein. Examples of peptides recognized by the binding entities of the invention include peptides having any one of SEQ ID NO: 89, 102, 110, 112, 127, 177, 182, 209, 474-488 or 489. Other examples of peptides recognized by the binding entities of the invention include peptides having any one of SEQ ID NO: 173, 185, 192, 196, 200, 490-491 or 492.

25   The binding characteristics of the binding entity can further differentiate between a phosphorylated peptide having any one of SEQ ID NO: 298, 301-324,326-347, 349-400, 402-410, 412-473, 571-643 or 644, and a non-phosphorylated peptide that differs from the phosphorylated peptide by substitution of Ser for the pSer or substitution of a Thr for the pThr. In some

30   embodiments, the phosphorylated peptide recognized by the binding entity can have any one of SEQ ID: 298, 320, 324, 350, 351, 366, 388, 394, 398, 402, 418, 464, 571-595 or 596. In other embodiments, the phosphorylated peptide recognized by the binding entity can have any one of SEQ ID: 301, 310, 317, 322, 344, 352, 371, 406, 597-599 or 600. For example, the phosphorylated

peptide recognized by the binding entity can have SEQ ID NO:298. Alternatively, the phosphorylated peptide recognized by the binding entity can have SEQ ID NO:313 or 314. Moreover, the phosphorylated peptide recognized by the binding entity can have SEQ ID NO:361 or 362.

5   The invention also provides a method for characterizing substrate specificities of kinases that includes: contacting each peptide pool in at least two test sets of peptide pools with ATP and a kinase; quantifying the amount of phosphorylation in each peptide pool; and comparing the amount of phosphorylation in each peptide pool with the amount of phosphorylation in at

10   least one other peptide pool. Test sets like those described above can be used in the methods of the invention. Comparison of the amount of phosphorylation in different peptide pools of a test set allows calculation of the preferences of the kinase for each query residue, which differs between those pools. By testing multiple test sets (for example, by using a superset described herein), a position

15   specific scoring matrix (PSSM) can be derived, which reflects the amino acid preferences of the kinase at positions around the phosphorylation position.

         The methods of the invention are flexible. For example, the same sets of degenerate peptides can be used to characterize many different kinases from every one of the millions of different biological species and an almost unlimited

20   range of mutant kinases derived from each such kinase. Flexibility is also present in the type of phosphorylation sites characterized by the methods of the invention and in the number of query positions and residue types are explored. Moreover, the methods of the invention can also be modulated so that different residues at a single position are tested, or the same residues are tested at different

25   positions. More than 500 peptide pools have been synthesized in more than 40 test sets, belonging to more than 6 supersets.

         The invention further provides a computer readable medium that includes computer-executable instructions, wherein the computer-executable instructions comprise conversion of input data into quantitative values specifying a

30   preference value for each of a plurality of amino acids at each defined position in a substrate peptide for a kinase, wherein: the input data comprises sequence and phosphorylation data for a test set of peptides comprising at least two peptide pools, wherein every peptide in each of the peptide pools comprises one phosphorylatable amino acid position, and one query amino acid position,

wherein: each peptide of every peptide pool has an identical phosphorylatable amino acid that can be phosphorylated by a kinase at the phosphorylatable amino acid position; the query amino acid position is at the defined position relative to the phosphorylatable amino acid position within every peptide of every peptide

5 pool but a query amino acid's identity at the query amino acid position is systematically varied from one peptide pool to the next peptide pool within the test set of peptide pools; a preference value for a particular amino acid at the defined position is substantially determined from the amount of phosphorylation of the peptide pool wherein that particular amino acid is the query residue and

10 the query position is located at the defined position.

        The invention also provides a method for visual display of amino acid or nucleotide sequence preferences comprising a series of stacks of single letter symbols for amino acids or nucleotides, wherein each stack represents a position in a peptide or a nucleic acid sequence; each symbol's height is proportional to

15 the absolute value of a quantitative parameter that is positive for favored amino acids or nucleotides and negative for disfavored amino acids or nucleotides; each symbol's position within the stack is sorted from bottom to top in ascending value by the quantitative parameter.

        In another embodiment, the invention provides a computer readable

20 medium having computer-executable instructions for performing a method of visually displaying amino acid or nucleotide sequence preferences, the method comprising: representing a position in a peptide or a nucleic acid sequence with a stack of single letter symbols for amino acids or nucleotides; and displaying a linear array of one or more stacks of letter symbols wherein each letter symbol's

25 height is proportional to the absolute value of a quantitative parameter that is positive for favored amino acids or nucleotides and negative for disfavored amino acids or nucleotides and wherein each letter symbol's position within the stack is sorted from bottom to top in ascending order by the value of the quantitative parameter.

30   The result of the graphic methods of the invention is a PSSM Logo, which is a novel graphical format for conveying the specificity information in a PSSM. It is particularly efficient in conveying both information on the preferred residues and the disfavored residues, which act in concert to determine the specificity of the kinase.

The present invention provides detailed information on the types of sites
and amino acid sequences that are recognized and phosphorylated by a kinase,
thereby permitting accurate prediction of which peptide sequences in the human
proteome can be phosphorylated by a particular kinase. Hence, computer

5   programs have been used to scan known well-defined human genes (15323).
Approximately 1900 human gene products were thereby identified that had at
least one Ser/Thr residue that predicted to be phosphorylated by protein kinase C
(PKC) using a high stringency prediction criterion (better than 0.2 percentile).
The validity of the PSSM derived results with supersets of peptides has been

10   extensively validated by demonstrating an excellent correlation between peptides
predicted to be phosphorylated *in vitro* by a kinase and those that are
phosphorylated *in vitro* by that kinase. Moreover, the biological relevance of the
in vitro phosphorylation is supported by comparison of sites identified with a
literature search defining sites phosphorylated *in vivo*.

15

Brief Description of the Figures

FIG. 1 provides examples of two test sets of peptide pools and results
obtained with PKC-theta using the methods of the invention.

FIG. 2 shows a superset of test sets designed for analysis of PKC

20   specificity from P-4 to P+3.

FIG. 3 provides counts per minute for *in vitro* phosphorylation by PKC-
theta of a superset of peptide pools designed for analysis of PKC specificity from
P-4 to P+3 for peptide pools shown in FIG 2.

FIG. 4 provides Ratio-to-Mean values for different amino acid residues at

25   different positions when using PKC-theta for peptide pools shown in FIG 2.

FIG. 5 provides a position-specific scoring matrix for PKC-theta using
the Log_2 Score for peptide pools shown in FIG 2.

FIG. 6 provides sequences of a superset of degenerate peptides designed
to extend analysis of PKC specificity.

30   FIG. 7 provides a position-specific scoring matrix for extended positions
using PKC-theta for peptide pools shown in FIG 6.

FIG. 8 illustrates the differences between the previously available
Sequence Logo for PKC (left) and a PSSM Logo of the invention for PKC-theta
(right).

FIG. 9 illustrates a validation study testing our predictions for PKC-theta and the previously available Scansite prediction for PKC-delta against results for PKC-delta. Each point on a given panel is a different peptide. The x-axis indicates a percentile prediction for phosphorylation of the peptide by PKC-theta

5   by our PSSM using data from P-4 to P+3 (panel A); by our PSSM using data from P-7 to P+6; and from Scansite for PKC-delta. The y-axis indicates phosphorylation of the peptide by PKC-delta expressed as percentage of phosphorylation of the best peptide.. Dashed lines indicate a reasonable thresholds for positive vs negative phosphorylation (at a value of 10%), and a

10   reasonable threshold for positive vs negative prediction (1st percentile). The curved line is an approximation of where points would be found for an optimal prediction. The results indicate that the predictions made according to the invention are valid and are better than the previously available Scansite method.

FIG. 10 compares the sensitivity and specificity of the present methods

15   with those provided by a previously available Scansite method using PKC-delta as the kinase.

FIG. 11 illustrates validation of the PKC-theta PSSM with a second set of proteomic peptides that were chosen for synthesis/testing based on prior knowledge of PSSM percentiles. Panel A shows results for individual peptides.

20   Panel B shows average results for groups of peptides grouped by PSSM percentile predictions.

FIG. 12 illustrates core sequences of a superset of test sets with 1 anchor position, represented by the formula d??R??S????d. Because of the importance of 'R' at P-3 to many basophilic kinases, these test sets are particularly useful for

25   such basophilic kinases.

FIG. 13 illustrates PSSM Logo for results of analysis of the kinase AKT1 with the d??R??S????d superset.

FIG. 14 illustrates proposed abundances of residues for use in degenerate positions. Also illustrated are hydrophobicity scores for each residue that has

30   been used in the invention to score hydrophobicity of peptides/sequences.

FIG. 15 shows detection of specific phosphorylation of SHP-1 by Western blot analysis using a pPKC antibody wherein the phosphorylation is augmented through stimulation by the T-cell receptor.

FIG. 16 provides a chart showing that scores derived from different test

sets tested at different times are reproducible and scores extrapolated for

untested residues can be adequately predicted.

FIG. 17 provides a graph of the data provided in FIG. 16, illustrating that

5·   scores derived from different test sets tested at different times are reproducible.

FIG. 18 illustrates how a peptide can be scored using data derived by the

methods of the invention.

FIG. 19 shows the distribution of scores observed when all Ser/Thr

containing sites in 15651 human proteins were scored with the PKC-theta PSSM

10   and shows the cutoffs for scores corresponding to particular low percentile

scores.

FIG. 20 illustrates that the PKC site prediction algorithm provided by the

invention correctly predicts previously known sites in the MARCKS protein.

FIG. 21 shows the high similarity in specificity between novel and

15   classical PKC isoforms, but atypical PKC differs more and great divergence seen

with AKT1 and PKA. Values shown are the Pearson correlation coefficients

derived from comparison of phosphorylation of panels of peptides by the kinase

pair indicated.

FIG. 22 illustrates the differences between PSSM Logos of different

20   kinases analyzed with the same peptide supersets.

FIG. 23 illustrates validation studies that demonstrate that the predictions

made for PKC-zeta are valid and are better predictions for PKC-zeta than for

PKC-delta.

FIG. 24 illustrates scoring changes in peptides that are less

25   phosphorylated by PKC-zeta than by PKC-delta.

FIG. 25 illustrates position-specific residue preferences for PKA and

PKG determined using the PKC superset.

FIG. 26 illustrates the differences between PSSM Logos of different

mutant kinases derived from PKC-theta analyzed with the same peptide

30   supersets. A PSSM Logo for wild type kinase analyzed using low levels of ATP

is shown in the lower right corner.

FIG. 27 illustrates the detailed changes in amino acid preferences

observed with PKC-theta mutant constructs and with altered kinase assay

conditions.

FIG. 28 illustrates that details of residue references for PKC-theta depend on the choices made for anchor and phosphorylation residues in the test sets used.

FIG. 29 illustrates results for ROK-alpha with test sets based on the ??R??T???? peptide set with only 4 query residues.

FIG. 30 illustrates details of the R-Pair Anchor optimization set.

FIG. 31 illustrates results for analysis of PKA with the R-Pair set shown in FIG. 30.

FIG. 32: shows that the R-Pair set reveals positions associated with the strongest preference for arginine (R).

FIG. 33 shows detection of specific phosphorylation of LIMK-2 by Western blot with the pPKC antibody which is augmented following stimulation by the T-cell receptor.

FIG. 34 shows detection of phosphorylation of MLK3 by Western blot with the pPKC antibody.

FIG. 35 is a diagram of a computerized system in conjunction with which embodiments of the invention may be implemented.

FIG. 36 shows RF-pair analysis for PKC-theta where the position of the arginine (R) and phenylalanine (F) residues is varied in a peptide having the sequence ddddddddSFddd, where "d" is a degenerate position in which either of the arginine or phenylalanine residues can be placed. Each peptide consisted of an N-terminal linker having a biotin-dansylated lysine and a glycine (BZG) followed by a 13 residue insert. The phosphorylation reactions were performed as described herein using PKC-theta as the kinase.

FIG. 37A-B shows average position-specific preferences of PKC-theta determined by the RF-pair (FIG. 37A) and R-pair (FIG. 37B) sets of peptides (see also FIGs. 30-32 and 36).

FIG. 38A-B illustrates that there is more than one strongly preferred RF-pair peptide for PKC-theta. FIG. 38B provides the structures of peptides (where "d" is a degenerate position) and their corresponding ratio-to-mean values with log2 score.

FIG. 39A-B provides an analysis of phosphorylation by the kinase PAK using an R-pair set of peptides. FIG. 39A is a chart showing how phosphorylation by PAK varies as the positions of the first and second arginine

residues are varied within the peptide set. FIG. 39B provides a graph of the Log2 score for arginine at various positions within a peptidyl sequence.

FIG. 40A-B provides an analysis of phosphorylation by the kinase PAK using an RF-pair set of peptides. FIG. 40A is a chart showing how

5   phosphorylation by PAK varies as the positions of the arginine and phenylalanine residues are varied within the peptide set. FIG. 40B provides a graph of the Log2 score for arginine (diamond symbols) and phenylalanine (square symbols) at various positions within a peptidyl sequence.

FIG. 41A-C provides an analysis of which arginine positions are favored

10   for phosphorylation by the kinase PAK using "diverse basic proteomic set" of peptides whose sequences are provided in Table 9. FIG. 41A shows the procedure for a chi-square analysis to determine whether arginine at position P-3 (relative to a phosphorylation site) contributes to phosphorylation of the 16 positively phosphorylated peptides. FIG. 41B provides the relative

15   phosphorylation of 16 peptides from the diverse basic proteomic set of peptides that have arginine at P-2 relative to the phosphorylated S or T. FIG. 41C shows the p-values for analysis of R at all positions between P-6 and P+3; the results demonstrate that R at P-2 is unique in its importance.

FIG. 42 shows that pPKC antibody binding requires the SHP-1 residue

20   S591 and that constitutively active PKC-theta (PKC-theta CA) can promote phosphorylation of the S591 residue. In the absence of the S591 residue (when using a S591A mutant), no phosphorylation by PKC-theta is detected.

FIG. 43A-B show that SHP-1 S591 is phosphorylated in T-cells in response to CD3/28 or PMA. Constructs with wild type or S591A mutant SHP-

25   1 sequences fused to GFP sequences were transfected into JURKAT or mouse thymocyte cells and SHP-1 phosphorylation was detected by western blot using an antibody specific for the phosphorylated SHP-1 S591 site (the "anti-S591 antibody"). As shown in FIG. 43A, the presence of serine at position 591 in SHP-1 is needed for phosphorylation. When alanine is present at position 591,

30   no phosphorylation is detected with the anti-pS591 antibody. FIG. 43B shows that T cell activation (using CD3/28 antibodies or PMA) in either the JURKAT cell line or in a mouse thymocyte preparation stimulates phosphorylation of the S591 residue of SHP-1.

FIG. 44 shows that PKC inhibitors BIM I and BIM III interfere with phosphorylation of SHP-1 at the S591 position.

FIG. 45A-D show that staining by anti-pS591 antibody is specific for SHP-1 Ser-591. No staining is observed when the S591A mutant of SHP-1 is

5   expressed (FIG. 45B).

FIG. 46A-C shows that phosphorylation of SHP-1 S591 inhibits nuclear localization of SHP-1.

Detailed Description of the Invention

10   The invention relates to determination of the specificity of protein kinases, to visual representation of specificity of kinases, to prediction of sites on sequenced proteins that are most likely to be phosphorylated by each kinase studied, to validation that peptides corresponding to those predicted sites are indeed phosphorylated *in vitro* by each kinase studied, and to validation of

15   phosphorylation of those sites *in vivo*.

The term "kinase" (or "protein kinase") as used herein is intended to include all enzymes that add a phosphate group to an amino acid residue within a protein or peptide. Kinases that may be used in the methods of the invention include protein-serine/threonine specific protein kinases, protein-tyrosine

20   specific kinases and dual-specificity kinase. Other kinases that can be used in the method of the invention include protein-cysteine specific kinases, protein-histidine specific kinases, protein-lysine specific kinases, protein-aspartic acid specific kinases and protein-glutamic acid specific kinases.

A kinase used in the method of the invention can be a wild type or

25   mutant kinase. The kinases employed can be purified native kinases, for example, a kinase purified from its native biological source. Kinases employed can be from a variety of species. Some kinases that can be employed are commercially available (e.g., protein kinase A from Sigma Chemical Co.). Alternatively, a kinase used in the method of the invention can be a kinase

30   produced by creation of a nucleic acid construct and preparing the protein product expressed *in vitro* or in whole cells (i.e., a "recombinantly produced kinase"). Many kinases have been molecularly cloned and characterized and thus can be expressed recombinantly by standard techniques. Hence, any recombinantly produced kinase that retains its kinase function can be used in the

methods of the invention. If the recombinant kinase to be examined is a eukaryotic kinase, it is generally preferable that the kinase be recombinantly expressed in a eukaryotic expression system to ensure proper post-translational modification of the protein kinase. Many eukaryotic expression systems (e.g.,

5   baculovirus and yeast expression systems) are known in the art and standard procedures can be used to express a kinase recombinantly. A recombinantly produced kinase can also be a fusion protein (i.e., composed of the kinase and a second protein or peptide) as long as the fusion protein retains the catalytic activity of the non-fused form of the kinase. Furthermore, the term "kinase" is

10   intended to include portions of native protein kinases that retain catalytic activity. For example, a subunit of a multi-subunit kinase that contains the catalytic domain of the kinase can be used in the methods of the invention.

One of skill in the art frequently uses a formula such as the following (I) to represent the amino acid positions within a peptidyl site that may be

15   phosphorylated by a kinase:

$$(P-4) - (P-3) - (P-2) - (P-1) - P0 - (P+1) - (P+2) - (P+3) - (P+4) \qquad I$$
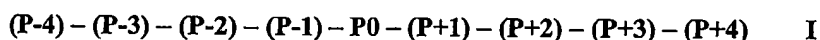
where P0 is the phosphorylated position, P-1 is the amino acid position immediately to the N-terminal side of P0, P+1 is the amino acid position

20   immediately to the C-terminal side of P0, P-2 is the amino acid position that is two residues from P0 on the N-terminal side of P0, etc. This terminology will be used herein as a general description of a kinase phosphorylation site and the variables P-4, P-3 etc. will be used to refer to a particular amino acid position within a kinase phosphorylation site.

25   In general, key positions that determine kinase specificity are within about four amino acids of the phosphorylated amino acid. However, positions farther than four positions from the phosphorylation site can influence the specificity of a kinase and can be characterized by the methods of the invention.

When one or more positions of a particular peptidyl sequence are

30   determined, a one letter amino acid symbol may be used herein to indicate what amino acid is present at that determined position. The standard three-letter and one-letter abbreviations for amino acids provided in Table 1 are used throughout the application.

TABLE 1

Amino acid	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Aspartic acid	Asp	D
Asparagine	Asn	N
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

The P0 position is the position that can be phosphorylated (the "phosphorylatable position") and is generally either a serine (S), threonine (T) or

5   a tyrosine (Y) for human kinases. Hence, specific peptidyl sequences generally discussed herein will often have S, T or Y at the P0 position. When any of a defined set of amino acids is present at a given position, for example, when a degenerate mixture of amino acids is used during synthesis of a peptide at that position, a lower case "d" is used herein to represent the degeneracy of that

10   position. To represent peptides in which a residue is phosphorylated, a lower case 'p' is often used herein before the residue abbreviation; thus, pS or pSer represents a phosphorylated serine residue, pT or pThr represents a phosphorylated threonine, and pY or pTyr represents a phosphorylated tyrosine.

15   **Design of single peptide test sets:**

The invention provides for determination of the specificity of protein kinases by synthesis of test sets (and supersets) of peptides, subjecting the test sets (or supersets) to phosphorylation by a kinase of interest, and quantifying and analyzing the results.

Two simplified embodiments shown in FIG. 1 are used as examples of
the methods provided herein. FIG. 1A shows one test set of peptide pools (a
"P+1" test set) and FIG. 1B shows a second test set (a "P+2" test set). As used
herein, the name of a test set generally identifies which position is being

5   systematically varied (i.e., which position is the "query" position. Each peptide
of the two test sets illustrated in FIG. 1 has a "core" sequence comprised of
eleven amino acid residues. The term "core" is used to refer to amino acid
sequences that play a key role in determining kinase specificity and is used to
distinguish such key amino acids from N-terminal or C-terminal residues that are

10   incorporated to provide functions unrelated to determination of specificity (such
as for capture of the peptide onto a solid support or for quantification).

Four different types of amino acid positions can occupy the core
positions in each of these peptides, as well as the other peptides described herein.
These different types of amino acid positions are described below.

15   1) A phosphorylatable amino acid position is a position occupied by an
amino acid to which a phosphate group can be added by a kinase. In eukaryotes
S, T, and Y are the primary phosphorylatable residues. However, in other
species residues such as histidine are also subject to phosphorylation. This
residue occupies the P0 position in each peptide pool in a test set. Hyphens (-)

20   may be used herein around the amino symbol in the P0 position (e.g., -S-) to
visually highlight this position. Note that the position of other types of amino
acid position in the core sequence are fixed relative to this P0 phosphorylatable
position in for all peptide pools in a given test set, and that each amino acid
position is expressed relative to the P0 position.

25   2) An anchor amino acid position is a position in addition to the
phosphorylatable amino acid position having a determined amino acid that does
NOT vary from one peptide pool to another in the test set. More than one anchor
amino acid position can be present in a test set. The location of the anchor
amino acid positions and identity of the anchor amino acids at each anchor

30   position are identical for *all* peptides pools in the test set. For example in the
P+1 set shown in FIG. 1A, there is one anchor amino acid: an arginine (R) at
position P-3. In the P+2 set, there are two anchor amino acids: an arginine (R)
at P-3, and a phenylalanine (F) at P+1. The function of the anchor amino acid
positions is to provide sufficient favorable interaction between substrate and

kinase to permit measurable phosphorylation of each peptide pool. An anchor amino acid is represented by a single letter amino acid code for the amino acid in that anchor position.

3) A query amino acid position (or a varied position) is a position that is
5   being tested for its effect upon substrate phosphorylation. The symbol "?" is often used herein as a symbol for identifying the query position. Unlike anchor amino acid positions, there is generally only a single query amino acid position within all peptide pools of a test set. In general, a query amino acid is determined (i.e., not degenerate) for a particular peptide pool. However, the
10   query amino acid at that query position is systematically varied from peptide pool to peptide pool within a test set of peptides. Hence, in contrast to the anchor positions, the query or varied position is occupied by different residues within the different peptide pools of a test set. The query or varied position is boxed in FIG. 1. The function of the query or varied positions is to allow
15   assessment of the contribution of different amino acids to kinase specificity by determining how each of the different tested amino acids influences the amount of phosphorylation.

4) A degenerate position contains an undetermined amino acid selected from a defined mixture of amino acids. More than one degenerate position is
20   typically present in a test set of peptide pools. For any given peptide pool in a test set, all core positions that are not anchor, phosphorylatable or query positions are degenerate positions. Thus, the presence of one or more degenerate positions means that each peptide pool in a test set of peptides is actually a complex mixture (or "library" of distinct peptides). Although each peptide pool
25   consists of many individual peptides, that peptide pool is often referred to herein as a "peptide," in keeping with common usage in the literature. Measuring phosphorylation of each such peptide pool assures that the assay reflects the average behavior of a large number of individual sequences. The symbol "d" is used herein as symbol of a degenerate position in the test sets of peptide pools
30   provided herein.

In some embodiments, the query position is not adjacent to an anchor position within the test sets provided herein. In other embodiments, the query position is not adjacent to the phosphorylatable position.

FIG. 1 illustrates the symbolic representation of two test sets of peptides
designed for analysis of PKC specificity, and the corresponding peptides pools
synthesized for those test sets. The formula **ddddRdd-S-?-dd** describes the P+1
test set of peptides shown in FIG. 1, where serine is in the P0 position, the query

5   position is P+1, arginine is the anchor amino acid chosen for an anchor position
at P-3 and the remaining amino positions are degenerate. Similarly, the formula
ddddRdd-S-F-?-d describes the P+2 test set of peptides shown in FIG. 1,
where: serine is in the P0 position, the query position is P+2; arginine is the
anchor amino acid chosen for an anchor position at P-3 ; phenylalanine is an

10   anchor amino acid chosen for a second anchor position at P+1; and the
remaining amino acid positions are degenerate (d).

Each test set in the embodiments shown in FIG. 1 consists of 13 peptide
pools. The residue present at the query position in each peptide pool in a test set
is systematically varied. However, the fixed anchor positions within all peptides

15   pools of the test set provide at least a minimal level of kinase recognition and
phosphorylation for each peptide in the test set. At the remaining core positions,
an amino acid selected from a degenerate mixture of amino acids is used.

Analysis of kinase specificity by phosphorylation of test sets

Determination of kinase specificity is made by phosphorylating the test

20   sets of peptides with a kinase of interest. Methods of the invention for
determining the substrate specificity of a kinase generally involve contacting
each peptide pool in at least one test set of peptide pools with a kinase and a γ-
labeled ATP, quantifying the amount of label incorporated into each peptide
pool, and comparing the quantity of label incorporated into a peptide pool with

25   the quantity of label incorporated into at least one other peptide pool.

Hence, a test set of peptides is synthesized, for example, the P+1 test set
having the thirteen sequences shown in FIG. 1 panel A. The synthesized peptide
pools in the test set are reconstituted to standardized concentrations, and
replicate samples of the peptide pools are contacted with a kinase under assay

30   conditions that permit phosphorylation at the P0 position. The amount of
phosphorylation of each peptide pool can be determined, for example, by
observing the radioactivity incorporated into the peptide pool after using γ^{32}P-
ATP as a donor of the phosphate group during the phosphorylation assay.

FIG. 1 panel A provides results of such a phosphorylation assay for the P+1 test set of peptides. The "raw data" are measured as counts per minute (cpm). As shown in FIG. 1, marked variation exists in the amount of phosphorylation present in different peptide pools of a test set, reflecting

5   important contributions of the single residue by which they differ. Furthermore, the SEMs (standard error of the mean of replicate values) are small, indicative of good assay agreement between replicates.

In some embodiments, the determination of residue preference is made by comparing the cpm incorporated into each peptide, with the geometric mean

10   cpm incorporated for all the peptides in the set. That ratio is shown in FIG. 1 within the column labeled 'Ratio-to-Mean.' The Ratio-to-Mean is also referred to herein as residue preference. A Ratio-to-Mean greater than 1.0 indicates that the selected query residue in the corresponding peptide is preferred by the kinase over the other types of query residues tested. For example, a Ratio-to-Mean of

15   2.9 was observed for 'F' in the P+1 test set, indicating that phenylalanine at P+1 is highly preferred by the kinase used for this assay (PKC-theta). A ratio less than 1.0 indicates that the selected query residue in the corresponding peptide pool is disfavored compared to the other residues tested. For example, a ratio of 0.4 was obtained for 'D' in the P+1 test set, indicating that aspartic acid at P+1 is

20   disfavored by the kinase used for this assay. To visually emphasize the preferred residues, the log scores in FIG. 1 for favored residues with residue preferences greater than 1.5 are in bold and underlined. In contrast, data relating to disfavored residues are bold without inderlining, indicating that the residue preference is less than 0.67 (i.e. 1.0 divided by 1.5).

25   A value called 'Log Score' (also called Log2 Score) was calculated for each residue by determining the log (base 2) of the Ratio-to-Mean. As a result of this mathematical transformation, favored residues have a positive score, and disfavored residues have a negative score. This score obviously differs depending on the position of the residue in the peptide (compare the P+1 test set

30   in FIG. 1A with the P+2 test set in FIG. 1B). Hence, each value represents a position-specific score for a particular amino acid residue. As indicated in FIG. 1 panel A, arginine, lysine, phenylalanine and leucine are preferred residues at the P+1 position for the kinase tested (PKC-theta). In contrast, aspartic acid,

asparagine, proline, glycine and alanine are disfavored at the P+1 position for the
kinase tested (PKC-theta).

The invention provides computer-executable instructions for performing
the calculations described above. One preferred embodiment uses software tools

5   enabled by use of a spreadsheet application such as Microsoft Excel running on
operating system such as Windows 2000 on a hardware platform such as a Dell
Latitude using a microprocessor such as an Intel Pentium chip. For example, a
spreadsheet is customized for a given superset of test peptides; manipulation of
that data is provided by formulas embedded in that spreadsheets. Output of

10   counts per minute from a TopCount NXT Microplate Scintillation and
Luminescence Counter in a 96 well plate format were inputted into the
spreadsheet. The results are displayed to the user in the spreadsheet; FIG. 3,
FIG. 4. and FIG. 5 are screen captures from such a spreadsheet. In one
embodiment additional processing of data is provided by automation of

15   additional functions in the spreadsheet using the language Visual Basic for
Applications, which is embedded in the Excel application; in other embodiments
additional automation is provided by software objects exposed by the Excel
interface and manipulated by software external to Excel, such as Microsoft
Visual Basic. This embodiment uses this same computational infrastructure for

20   performing the manipulations described in Example 3.

Thus, the invention provides a computer readable medium having
computer-executable instructions for determining quantitative values describing
the preference of a kinase for a defined amino acid at a defined substrate position
wherein the input data comprises experimental data on phosphorylation of a test

25   set of peptides comprising at least two peptide pools, wherein every peptide in
each of the peptide pools comprises one phosphorylatable amino acid position,
one query amino acid position, wherein each peptide of every peptide pool has
an identical phosphorylatable amino acid that can be phosphorylated by a kinase
at the phosphorylatable amino acid position and the query amino acid position is

30   at a defined position relative to the phosphorylatable amino acid position within
every peptide of every peptide pool but a query amino acid's identity at the
query amino acid position is systematically varied from one peptide pool to the
next peptide pool within the test set of peptide pools

Supersets constructed from multiple test sets

The test sets illustrated in FIG. 1 provide information on positions P+1 and P+2, based on the location of the query position relative to the phosphorylatable anchor residue. In general, all positions within a test substrate

5   can separately be made into query positions by constructing a test set of peptides for each query position. Hence, one of skill in the art can make, for example, P-7, P-6, P-5, P-4, P-3, P-2, P-1, P+1, P+2, P+3, P+4, P+5, P+6 and P+7 test sets of peptides and systematically vary the type of amino acid at each of these query positions. Such a large of test sets of peptide pools with query residues at

10 substantially all the different positions is referred to as a superset. In some embodiments, each position close to the phosphorylation site (P0) will be a query position and the appropriate test sets of peptides within the superset will be made and tested to ascertain which amino acid is preferred by the kinase at those query positions. FIG. 2 shows such a superset of test sets of peptides

15 designed and synthesized to test the specificity of PKC and related kinases at all query positions from P-4 to P+3. This superset includes the two test sets shown in FIG. 1 together with six other test sets.

Such supersets are phosphorylated by a kinase of interest as described for the test sets above. FIG. 3 shows the raw data (cpm) obtained for a

20 representative experiment testing PKC-theta on the superset shown in FIG. 2. FIG. 4 shows the Ratio-to-Mean for that data, calculated as described above. FIG. 5 shows the Log (base 2) score for that data, calculated as described above. Taken together, the scores derived from analysis of a superset of peptides (e.g. FIG. 5) constitute a position-specific scoring matrix (PSSM) describing the

25 residue preference of the selected kinase at different positions around the phosphorylation site.

A reduced set of amino acid residues can be used in the query position of the test sets of peptides. Experimental data obtained for such reduced sets of query amino acids do not provide information for all naturally occurring

30 residues. In some embodiments, data that is not obtained experimentally can be estimated from existing data. For example, the lower boxed region shown in FIG. 5 provides extrapolated data for residues that were not tested, but that have similar physicochemical properties to the peptides tested. Thus, in this case data for glutamic acid (E) was inferred from aspartic acid (D), data for isoleucine (I),

methionine (M) and valine (V) was inferred from leucine (L), data for tyrosine

(Y) was inferred from phenylalanine (F). Where cysteine was excluded from the

residues analyzed, a score for cysteine was likewise created from scores for other

residues. Such extrapolation can be accomplished in a variety ways, for

5   example, by assigning a score of zero, or assigning the score corresponding to

other residues such as alanine. The accuracy of these extrapolated scores can

then be tested as described below (Example 2).

        The method of the invention is flexible so that greater or lesser numbers

of test sets can be included for testing as many positions as desired. For

10   example, FIG. 6 lists the sequences of a superset of peptide pools designed to

extend the analysis of PKC specificity to include positions P-7 through P-5 and

P+4 thru P+6. FIG. 7 shows an extended position-specific scoring matrix for

positions P-7 through P-5 and P+4 through P+6 derived from testing PKC-theta

with the test sets shown in FIG. 6. Taken together, the scores from FIG. 5 and

15   FIG. 7 provide a position-specific scoring matrix for PKC-theta for positions P-7

to P+6. The ability to combine results from different sets and different

experiments is a convenient aspect of the invention.

Visual representation of kinase specificity

        An efficient strategy for visual representation of specificity information

20   is important for conceptualizing and communicating findings on kinase

specificity. A previously described method for visualizing peptide specificity

data is via the Sequence Logo developed by Thomas Schneider (Schneider TD et

al. 1990. Nucleic Acids Res. 18:6097-6100). In that article, the method is

described as follows "The height of each letter is made proportional to its

25   frequency, and the letters are sorted so the most common one is on top. The

height of the entire stack is then adjusted to signify the information content of

the sequences at that position." This visualization method is illustrated on the

left side of FIG. 8 for a published Sequence Logo generated by the Schneider

method for protein Kinase C (PKC) (Kreegipuu A et al. 1998. FEBS Lett

30   430:45-50).

        The invention provides a new method for visualizing which amino acids

are preferred in the substrate of a kinase. This method involves use of a position

specific residue scoring matrix (PSSM) to generate a PSSM Logo. Each

position in a PSSM is represented in a PSSM Logo by a vertical stack of amino

acid residue single letter codes. The height of each code is made proportional to the absolute value of a Log Score, and the positions of the codes in the stack are sorted from bottom to top in ascending value by the quantitative parameter. An example of a PSSM Logo of the invention is provided on the right side of FIG.

5   8, which illustrates the results for analysis of PKC-theta with peptide pools shown in FIG. 2 and FIG. 6.   In the preferred embodiment, each single letter code is colored to indicate the physico-chemical properties of the corresponding residue; for example R, K, H could be blue to indicate basic, D, E red to indicate acidic, I, L, M, V, F, Y could be grey to indicate hydrophobic.

10   Two major differences exist between the previously available Sequence Logo and a PSSM Logo of the invention. The most fundamental difference between a Sequence Logo and a PSSM Logo is that the PSSM Logo visually emphasizes the residues that are disfavored by the kinase as well as the ones that are favored by the kinase. In contrast, the Sequence Logo only emphasizes the

15 residues that are favored. Such distinction is not a trivial distinction, but rather represents a fundamental difference in emphasis between the method of the invention and those of prior workers. In particular, the present methods accurately determine which amino acid residues are disfavored, which has not previously been emphasized and which can be a controlling factor in

20 determining kinase specificity (see below).

            A secondary difference between the previously available Sequence Logo and a PSSM Logo of the invention is in the parameters represented by the PSSM Logo versus those represented by the Sequence Logo. The Sequence Logo, as described by Schneider, is determined by a combination of the parameters

25 referred to as 'information content' of that position, and of the residue frequency. In contrast, in a preferred embodiment, the PSSM Logo reflects the log scores obtained by the methods of the invention, which are not interchangeable with residue frequency. In other embodiments, the parameter represented in the PSSM Logo is the log of the ratio of [residue

30 frequency]/[control residue frequency]. Hence, the PSSM Logo is distinct from the Sequence Logo.

            Note that use of a PSSM Logo is not restricted to findings of kinase specificity, but rather is generally useful for expressing results pertaining to amino acid residue preference. Thus, for example, results of other experimental

methods for determination of residue preference for peptide binding (rather than phosphorylation) can equally well be represented with a PSSM Logo. Moreover, nucleotide sequence preferences can also be represented using a PSSM Logo.

One embodiment uses software tools enabled by use of a spreadsheet
5   application such as Microsoft Excel running on operating system such as Windows 2000 on a hardware platform such as a Dell Latitude using a microprocessor such as an Intel Pentium chip. Software objects exposed by the Excel interface are manipulated by software external to Excel, such as Microsoft Visual Basic. Information in the spreadsheet for each substrate position consists
10   of paired columns, one comprising the residue code and one comprising the log2 scores. Rows in that pair of columns are sorted in descending order by log2 scores. That sorted information is converted into a file of commands using postscript programming language which instruct a postscript printer (such as Xerox Phaser 6200 printer) to create symbols of the appropriate size and
15   position in a column. Successive columns in the PSSM are processed similarly and the postscript code instructs the printer to move horizontally to position information on each successive substrate position into adjacent columns.

Thus, the invention provides a computer readable medium having computer-executable instructions for performing a method of visually displaying
20   amino acid or nucleotide sequence preferences, the method comprising: representing a position in a peptide or a nucleic acid sequence with a stack of single letter symbols for amino acids or nucleotides; and displaying one or more stacks of letters wherein each symbol's height is proportional to the absolute value of a quantitative parameter that is positive for favored amino acids or
25   nucleotides and negative for disfavored amino acids or nucleotides and wherein each symbol's position within the stack is sorted from bottom to top in ascending value by the quantitative parameter.

The invention also provides an overview of the hardware and the operating environment in conjunction with which embodiments of the invention
30   can be practiced. Figure 35 is a diagram of a computerized system in conjunction with which embodiments of the invention may be implemented. Thus, in one embodiment, computer 110 is operatively coupled to a monitor 112, a pointing device 114 and a keyboard 116. Computer 110 includes a central processing unit 118, random-access memory (RAM) 120, read-only memory

(ROM) 122, and one or more storage devices 124, such as a hard disk drive, a floppy disk drive, a compact disk read-only memory (CD-ROM), an optical disk drive, a tape cartridge drive or the like. RAM 120 and ROM 122 are collectively referred to as the memory of computer 110. The memory, hard drives, floppy

5   disks, etc., are types of computer-readable media. The computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for computer 110. The invention is not particularly limited to any type of computer 110.

         Monitor 112 permits the display of information for viewing by a user of

10   the computer. Pointing device 114 permits the control of the screen pointer provided by the graphical user interface of window-oriented operating systems such as the Microsoft Windows family of operating systems. Finally, keyboard 116 permits entry of textual information, including commands and data, into computer 110.

15   The computer 110 operates as a stand-alone computer system or operates in a networked environment using logical connections to one or more remote computers, such as remote computer 126 connected to computer 110 through network 128. The network 128 depicted in Figure 34 comprises, for example, a local-area network (LAN) or a wide-area network (WAN). Such networking

20   environments are common in offices, enterprise-wide computer networks, intranets, and the Internet.

         An example hardware and operating environment in conjunction with which embodiments of the invention can be practiced has been described.
Validation of the results obtained using the methods described

25   One of the principle uses for the methods of the invention is to predict sites of phosphorylation in proteins whose sequences are known but whose phosphorylation sites are unknown. The ability to correctly predict phosphorylation sites will depend on the correctness of the methods employed. If the values for residue preference in for a kinase are incorrect, then the

30   predictions are unlikely to be correct. As described herein a PSSM generated by the methods of the invention will generally provide better and more complete substrate specificity information than previously employed methods and predictions employed.

Rather surprisingly, systematic validation has not been reported for previously reported predictive algorithms, such as those proposed by U.S. Patent 6,004,757 to Cantley et al. For example, Nishikawa K et al. 1997. J Biol Chem 272:952-960 describes an approach for determining peptide specificity for PKC,

5   but the validation provided was limited to a showing that the optimal peptides predicted for two different kinases are preferentially phosphorylated by their respective kinases. No validation was provided that the sequence identified was the best sequence, or that good *in vitro* substrates can be identified by using the remainder of the information derived from the technique. While, Cantley and

10   co-workers also propose that the results of such predictions correlate with physiologically relevant sites, such assertions are based on a modest correlation with anecdotal results from the literature.

One approach to validating a substrate identification method can involve, for example, comparison of substrate sites predicted by the method with *in vitro*

15   phosphorylation results obtained using the selected kinase and peptides of known sequences. Such a systematic validation has been performed for the methods described herein. For example, a panel of seventy five peptides was synthesized, the phosphorylation observed for each peptide was experimentally measured, the amount of phosphorylation was quantified, the phosphorylation

20   results for each peptide were normalized to the phosphorylation observed with the best substrate tested and these amounts were compared with predictions made according to the invention and according to the procedures provided by others. These peptides are referred to herein as proteomic peptides because their sequences are chosen from proteins in the human proteome; unlike the test sets

25   employed herein, these peptides include no degenerate positions

Fairness of a validation strategy requires that the choice of test peptides not be unfairly biased by findings from the PSSM being validated. The choice of the peptides in Table 2 was not biased by information from the PSSM-based scoring illustrated herein because the peptides were chosen and synthesized

30   more than five months before the method was established. The dominant criteria for selection of the peptides was computerized scanning of human protein sequences amongst NCBI reference sequences (see website at ncbi.nlm.nih.gov/) to identify sites with an abundance of positively charged residues in positions P-

3 to P+3 relative to a potential P0 phosphorylation position (S or T), and with good diversity in the P-1 and P+1 positions.

The results of this analysis for phosphorylation are provided in Table 2. While the results provided in Table 2 show measured phosphorylation by PKC-

5   delta, the PKC-delta predictions made by the methods of the invention (shown in Table 2) were actually based upon data obtained by PKC-theta. In contrast, data generated by the methods of Cantley and co-workers was available for PKC-delta (Nishikawa K et al. 1997. J Biol Chem 272:952-960; and Scansite at scansite.mit.edu). Because the predictions from the present methods are based

10   on PKC-theta, which is distinct from PKC-delta but is the PKC isoform closest to PKC-delta, the comparison provided in Table 2 is biased in favor of the method provided by Cantley and co-workers. Despite this bias, the results demonstrate that predictions made by the methods of the invention are better than predictions made by the methods of Cantley and co-workers (Scansite).

15

Table 2: Validation of the Present Methods
Comparison of Present Method vs. Scansite Predictions

SEQ ID NO:	Sequence	Prediction (percentile)		Measured in vitro phosphorylation by PKC-delta
		Invention for PKC-theta	Scansite for PKC-delta	
1	HVRRRRGTFKRSKLRARD	0	0.26	100
2	KKKKRASFKRKSSKKG	0	0.01	76
3	NRKKKRTSFKRKA	0.1	0.05	66
4	KFARKSTRRSIRLPE	0.9	4.29	52
5	RQRKRKLSFRRRTDKD	0	0.35	42
6	PRLIRRGSKKRPAR	0	>5	40
7	RKIPKRPGSVHRTPSRQ	0.2	4.23	38
8	AARKKRISVKKKQEQ	0.2	0.04	35
9	QKKSRLRRRASQLKI	0.1	3.83	34
10	AQIVKRASLKRGKQ	0.5	0.03	32
11	KKKFRTPSFLKKSKK	0.4	1.52	25
12	KKKKKRFSFKKSFKL	0.2	0	24
13	WKGKRRSKARKKRK	2.5	>5	22
14	EYLERRASRRRAV	0.1	>5	20
15	RGFLRSASLGRRASFHLE	0	0.41	18
16	DGQKRKKSLRKKLD	0	>5	17
17	AGWRKKTSFRKPKED	0.2	0.75	17
18	KKRFSFKKSFKLSGFSFKKN	0.2	0.01	16
19	AGSFKRNSIKKIV	0.3	1.69	14
20	GAPPRRSSIRNAH	0.4	>5	13
21	KLAVGRHSFSRRSGV	0.5	>5	12

SEQ ID NO:	Sequence	Prediction (percentile)		Measured in vitro phosphorylation by PKC-delta
		Invention for PKC-theta	Scansite for PKC-delta	
22	LLKKRDSFRTPRDSKLE	2.5	2.51	12
23	QKRHARVTVKYDRRE	1.5	4.49	10
24	EKIKRSSLKKVDSLKK	1.5	0.02	10
25	EILSRRPSYRKILND	0.1	>5	9
26	ALRRPSLRREADD	0.2	>5	9
27	KKRKKKSSKSLAHA	2.7	0.02	8
28	KRPGKKGSNKRPGKR	4	0.48	8
29	RKNDRKKRYTVVGNP	>5	>5	8
30	KEVVRTDSLKGRRGR	1.5	>5	7
31	RKKRKKKSSKSLAHAGVALA	2.7	0.02	>5
32	KATTKKRTLRKNDRK	1.7	0.48	>5
33	QQKIRKYTMRRLLQE	0.5	>5	>5
34	EGGDRRASGRRK	2.1	>5	5
35	GLLDRKGSWKKLDDM	2.1	3.26	4
36	GENVLKKSMKSRVKG	5.2	>5	4
37	AYIERMNSIHRDLRA	3.1	>5	3
38	NYLRRRLSDSNFMAN	0.9	>5	3
39	LLGSGKVTDRKAL	>5	>5	3
40	NMEAKKLSKDRMKKY	>5	>5	3
41	FVHQASFKFGQGD	1.5	0.04	3
42	QPEGLRSLKKPDRKKR	>5	>5	3
43	AWVTVHEKKSSRKSEYL	4.2	2.95	3
44	VLAKKGTSKTPVPE	>5	2.43	2
45	VFREHQRSGSYHVRE	0.1	>5	2
46	GQAWGRQSPRRLED	>5	>5	2
47	ARIIGEKSFRRSVVG	2.7	0.69	2
48	AVNSRRRAGQKKK	5	>5	2
49	VQQLLRSSNRRLEQL	>5	>5	2
50	ENLRRVATDRRHLGH	0.8	>5	2
51	DLLGKKVSTKTLSEDD	>5	4.05	2
52	HKHSPEKRGSERKEG	>5	>5	2
53	AKNLKTLQKRDSFIG	>5	0.41	2
54	ENLRKVTTDKKSLAY	>5	0.01	2
55	DDMEHKTLKITDFG	1.5	>5	2
56	EARLGAASLKFGARD	>5	0.01	2
57	KNVVKLLSSRRTQDR	>5	4.49	2
58	RVKLGTLRRPEGP	>5	4.05	1
59	PVNKRSKYTMMK	4.1	0.18	1
60	LRRKHLGTLNFGGIR	>5	0	1
61	VDNILKKSNKKLEEL	5.3	>5	1
62	AVRDMRQTVAVGVIK	>5	0.84	1
63	QRQERIFSKRRGQDF	3.4	>5	1
64	ALRAPKPTLRYFTTERF	>5	0	1
65	IKVTHKATGKVMVMK	>5	>5	1
66	GFAKKIGSGQKTWTF	>5	0.15	1
67	AINSRETMFHKERFK	>5	>5	1
68	RGEGHKPSIAHRDFK	>5	>5	1
69	LALTARESSVRSGGAG	>5	0	1
70	HERKGSDKRGDNQ	4.1	>5	1

SEQ ID NO:	Sequence	Prediction (percentile)		Measured in vitro phosphorylation by PKC-delta
		Invention for PKC-theta	Scansite for PKC-delta	
71	RRRQKRRTGALVLSRGGKR	>5	>5	1
72	LTDPKEDPIYDEPEGLAPVPG	>5	>5	0
73	IDYYKKTTNGRLPVK	>5	>5	0
74	IDYYKKTSNGRLPVK	>5	>5	0
75	EEAEHKATKARLADK	>5	>5	0

Two steps are involved in the validation process: making the predictions, then assessing the predictions by comparison with measured values. When a PSSM is obtained by the methods of the invention, the calculation of a
5   prediction is straightforward, using the algorithms described herein (see, e.g., example 3).

Table 2 compares the present predictions with actual measurements of phosphorylation on validating peptides. The method of synthesis of the validating peptides was as described elsewhere in the application, and each
10   included an N-terminal linker sequence of biotinylated-Lys-dansylated-Lys-Pro-Pro-Gly (SEQ ID NO:231). The length of the remaining "core" of the validating peptides ranged from 12-21 residues with one to five S/T residues. *In vitro* phosphorylation of these validating peptides was measured in the manner described herein. Measurements were obtained by phosphorylation of the
15   validating peptides with PKC-delta at a peptide concentration of 10nM. *In vitro* phosphorylation results for the validating peptides were expressed as normalized values, namely as a percentage of phosphorylation of the best validating peptide substrate in the group. Hence, a higher value for the measured *in vitro* phosphorylation of a validating peptide indicated that the validating peptide was
20   phosphorylated to a greater extent than a validating peptide with a lower phosphorylation value.

Many of the peptides employed (Table 2) have multiple serine/threonine residues; the score for a peptide is determined by scoring each Ser/Thr in the peptide and the lowest (i.e. best) percentile for all residues that could be
25   phosphorylated was taken as the percentile for the peptide.

In addition to the measured value, Table 2 tabulates percentile prediction scores for the validating peptides where the prediction scores were obtained

either by the methods of the invention or by the methods of Cantley and co-workers. To obtain predictions made as described by Cantley et al, the sequence of the peptide was analyzed using Scansite (see website at scansite.mit.edu/). Scansite is a website made publicly available by L. Cantley and M. Yaffe to

5   predict best substrates based on data derived by the Cantley degenerate peptide strategy. By both the present methods and by the methods of Cantley, a lower positive prediction value indicated a stronger prediction that the peptide will be phosphorylated. Using the conventions of Scansite, predictive percentile scores greater than 5 were shown as >5.

10   As shown in Table 2, FIG. 9, and FIG. 10 , the methods of the invention are better predictors of which peptide sequence will be phosphorylated than are the methods provided by the prior art. For example, peptide SEQ ID NOs: 4, 7, 9 and 11 were highly phosphorylated by the *in vitro* validating assay but the Scansite methods predicted significantly poorer levels of phosphorylation than

15   did the methods of the invention. Similarly, peptide SEQ ID NOs:60, 64, 66 and 69 were poorly phosphorylated by the *in vitro* validating assay but the Scansite methods predicted significantly higher levels of phosphorylation than did the methods of the invention.

The predictive accuracies of the methods of the invention and those of

20   Cantley and co-workers (Scansite) are summarized in FIG. 9. FIG. 9 provides a correlation between the predicted percentile and the measured phosphorylation for each peptide. Results are shown for three different predictions: predictions of the invention based only on positions -4 to +3 for PKC-theta; predictions of the invention based on positions -7 to +6 of PKC-theta and the Scansite

25   prediction for PKC-delta. A curve has been overlaid on each of the three plots to indicate what the correlation might be expected to look like. Note that accurate predictions will have few peptides in the upper right (false negatives) or the extreme lower left (false positives). Inspection of FIG. 9 reveals that predictions made by using the methods of the present invention are both good, and that the

30   expansion from P-4/P+3 to P-7/P+6 gives modestly improved predictions. In contrast, the pattern observed with the Scansite prediction includes many more peptides that are located at positions far from the optimal correlation.

FIG. 10 tabulates the results obtained. As shown in FIG. 10, the methods of the invention have approximately 90% specificity and sensitivity while the

methods provided by Scansite have only 70% specificity and 45% sensitivity. Thus, the methods provided by the invention for predicting kinase specificity are better than this prior art approach for predicting PKC-delta specificity, even though the analysis was weighted in favor of the Cantley approach by using

5 PKC-delta, which was exactly the kinase that Cantley used, and only a close relative of the kinase used in the methods of the invention (PKC-theta).

Identification of peptides efficiently phosphorylated by PKC

A second strategy for validation of the PSSM derived from the methods described herein is to identify sequences represented in the human proteome that

10 have low percentiles derived from the PSSM, to synthesize peptides that have those sequences, and test the efficiency of phosphorylation of those peptides by the kinase of interest. FIG. 11 shows the results for such an analysis for 96 individual peptides. The results are shown for individual peptides (FIG. 11, panel A) or for groups of peptides aggregated by percentile prediction (FIG. 11,

15 panel B). As with the testing described above with prospectively chosen peptides, the percentile scores are highly predictive of phosphorylation by the relevant kinase.

The process of prediction and testing resulted in identification of many peptides predicted to be substrates for PKC-theta and demonstrated to be

20 substrates for PKC-theta (Table 3). A number of the sequences surrounding the most likely phosphorylation site have quite incomplete matches to the prototypic PKC substrate pattern [RK][RK]x[ST][hydrophobic][RK][RK]. Most of these peptides/sites have not previously been reported to be substrates for PKC in vivo or *in vitro*.

25

TABLE 3.

Identification of *in vitro* substrates of PKC-theta with further method validation

Sequence	SEQ ID NO	Locus-LinkID	Name	Measured in vitro phosphoryla-tion by PKC-theta	Prediction from PKC-theta
-AMSRSA-S-KRRSR-	168	7074	TIAM1	100	0.5
RTRSRRL-T-FRK---	169	1901	S1P1 receptor	100	0.0

Sequence	SEQ ID NO	Locus-LinkID	Name	Measured in vitro phosphoryla-tion by PKC-theta	Prediction from PKC-theta
--VKLRR-S-KKRTKR	170	1794	DOCK2	98	0.1
--RRGRRSTKKRRR	171	55672	FLJ20719	92	0.0
--VRRRRSQRISQR	172	25836	IDN3	86	0.0
RSGRRRGSQKS---	173	202	absent in melanoma 1	85	0.0
KKERRRNSINRN--	174	4542	myosin IF	83	0.0
-KKRRTKSSRRGV-	175	1612	DAP-kinase 1	80	0.1
---RRERSRSRRKQ	176	2305	forkhead (Drosophila)-like 16	66	0.1
-RRRRRRSRTFSR-	177	1196	CLK2	66	0.0
---RRRRSRTFSRS	178	1196	CLK2	65	0.0
-KRHYRKSVRSRS-	179	65125	WNK1	65	0.1
-FLRRSSSRRNRS-	180	9595	PSCDBP	65	0.1
TGERKRKSVRG---	181	6194	ribosomal protein S6	62	0.3
-TKKKRGSYRGGS-	182	9221	nucleolar phosphoprotein p130	61	0.6
-ARRSKRSRRRET-	183	23031	MAST3	55	0.1
----FRASSRSTTK	184	4863	NPAT	54	1.0
KKFKRRLSLTLR--	185	5128	PCTK2	51	0.1
-DFRRRRSFRRIA-	186	5734	prostaglandin E receptor 4	50	0.0
--LRRKSSTRHIHA	187	672	BRCA1	48	0.2
-ERGRRGSKKGSI-	188	695	BTK	44	0.1
GRRRRSRSKVK---	189	8899	serine/threonine-protein kinase PRP4	43	0.0
RRRRHTMDKDSR	190	65125	WNK1	40	0.1
---HKRNSVRLVIR	191	409	beta-arrestin2	38	0.5
GNRKGKSKKWRQ-	192	2870	GRK6	35	0.5
--PLRKSSLKKGGR	193	393	ARHGAP4	35	0.3
-KRRKRKSLQRHK-	194	1455	casein kinase I gamma 2	34	0.1
PGSSHRKTKK---	195	695	BTK	33	0.8
-RWKRRRSYSREH-	196	1198	CLK3	32	0.1
-ILRPSKSVKLRS-	197	26191	Lyp	32	0.6
--RRRRPTKSKGSK	198	65125	WNK1	28	0.0
-RGRRSRSRLRRR-	199	8899	serine/threonine-protein kinase PRP4	27	0.0
EQQRRALSFRQ---	200	5778	HePTP	26	1.0
-TQDRRKSLFKKI-	201	23031	MAST3	25	0.2
-VMKRKFSLRAAE-	202	6840	supervillin	24	0.6
-VRRSKKSKKKES-	203	23227	MAST4	24	0.3

Sequence	SEQ ID NO	Locus-LinkID	Name	Measured in vitro phosphoryla-tion by PKC-theta	Prediction from PKC-theta
RFSRRSSSWRIL--	204	4033	LRMP	22	0.6
-EGRRSRSRRYSG-	205	1105	CHD1	22	0.1
KSSRNSTSVKKK--	206	9934	GPR105	19	0.3
-SFRGHITRKKLK-	207	2596	gap-43	18	0.2
-VSRPRKSRKRVD-	208	25836	IDN3	17	0.2
DKEKSKGSLKRK--	209	5777	SHP-1	17	2.0
-PLRRRESMHVEQ-	210	6650	SOLH	16	0.1
RSRSYSRSRSR---	211	4820	NKTR	16	1.0
--VSRGSSLKILSK	212	7852	CXCR4	13	2.0
-RHSRSRSRHRLS-	213	8621	CDC2L5	13	0.8
-SRRRSPSYSRHS-	214	8621	CDC2L5 serine/threonine-protein kinase PRP4	13	0.3
-TKKRSKSRSKER-	215	8899		12	0.5
--SCRTSSRKRAGK	216	8915	BCL10	11	1.0

Considerations in design of test sets of peptides

      Design of each test set of peptides involves important decisions regarding: the choice of phosphorylatable residue, the choice of anchor

5   positions, the identity of residues at the anchor positions, the choice of the query positions, the identity of residues for the query positions and choice of positions and residue types for the degenerate positions. These considerations are discussed in more detail below.

      In most embodiments, one position is a residue that can be

10   phosphorylated (a phosphorylatable amino acid position), such as serine (S), threonine (T) or tyrosine (Y). As described above such a phosphorylatable position is referred to as "P0." The choice between S, T and Y is based on the known or inferred phosphorylation preference of the kinase(s) whose specificity is to be assessed. For example, protein kinase C (PKC) phosphorylates a serine

15   (S) more often than threonine (T). However, data obtained by the inventors indicates that Rho-kinase generally phosphorylates a threonine (T) and it has been previously determined that Lck generally phosphorylates a tyrosine (Y). Hence, one of skill in the art can use available information to assign the identity of the phosphorylatable amino acid. Alternatively, procedures like those

20   provided herein or other available procedures can be used to determine which residues are preferentially phosphorylated by a kinase of unknown specificity.

Selecting the number and identity of Anchor Positions.

Anchor positions in the peptides used in the present methods can be at any position within the sequence of a test peptide pool. In particular, anchor positions do not need to be contiguous (i.e. next) to each other in the present methods. Anchor positions need not be adjacent to the query amino acid position. Anchor positions also do not need to be adjacent to the phosphorylatable residue. For example, many of the test sets in the superset of peptides used for PKC analysis had anchor residues in the pattern Rxx-S-F (see FIG. 2) where the anchor residue arginine (R) was adjacent neither to the phosphorylatable residue serine (S) nor to the other anchor residue phenylalanine (F).

The number of anchor positions selected for a set of peptides can influence the amount of information obtained about the substrate. In general, if too many residues are anchored then the test set will be relatively insensitive to changes in the query residues. However, if too few residues are anchored then the average amount of phosphorylation in the set will be too low. Low levels of phosphorylation can lead to error-prone readings. For example, when there is a low level of phosphorylation, decreases in phosphorylation caused by disfavored query residues will generally be small and unreliable.

In most embodiments, one or two positions are assigned to be anchor positions. However, a larger number of anchor residues can be useful in some embodiments, particularly those designed for particular conditions. As illustrated herein some embodiments have two anchor positions. For example, two anchor residues were used for six of the eight test sets in a superset design for PKC analysis, i.e. R??-S-F?? (FIG. 2). As show herein, use of this superset provides a good characterization of the specificity of PKCs.

Supersets with one anchor position are also very useful. The utility of such a superset with one anchor position is illustrated by a superset consisting of 8 test sets with the symbolic representation d??R??S????d (FIG. 12). This d??R??S????d superset is an especially useful superset for initial characterization of kinases that may be basophilic, because many basophilic kinases have a strong preference for 'R' at the P-3 position.

FIG. 13 shows a PSSM Logo for analysis of the kinase AKT1 with this superset, which provides a good overview of the preferences of AKT1 at most

positions between P-5 and P+4. Because there is only one anchor residue, the counts per minute for this superset after phosphorylation are typically lower than with two suitable anchor positions. However, this superset can still provide an adequate "dynamic range" showing favored and disfavored residues (FIG. 13).

5   Data from this analysis provides an approximation of the specificity of AKT1. If more precise understanding is required, then a suitable second anchor position can be chosen from the results of this d??R??S????d set, and an additional superset(s) of test peptides can be synthesized with two anchor positions. One of skill in the art can envision other one-anchor sets that would be especially useful

10   such as d?????SP???d for proline-directed kinases, d?????SQ????d for 'SQ' directed kinases, and d?????SR???d for 'SR' directed kinases.

According to the invention, several principles for choosing a second anchor position from the results of a one anchor set such as d??R??S????d. In general, the second anchor is an amino acid that is strongly preferred by the

15   kinase of interest. In the case of AKT1, illustrated by FIG. 13, there are multiple such residues, for example, R at P-5, R at P-2, and F at P+1. In choosing between those, a secondary consideration is minimizing the number of other preferred residues at that position. Hence, a second anchor amino acid is selected as the most preferred of only a few preferred residues at that position.

20   Based on that criterion, a particularly good choice would be R at P-5. If one of skill in the art wishes to obtain more detailed information on which anchor residues to select, multiple second anchors can be chosen and supersets synthesized to test each anchor position.

It is also important to note that a superset based on no anchors, such as

25   d????S????d or d????Y????d can also be useful. Information derived by analysis with such a set could be particularly useful for choice of a second anchor (distinct from R at P-3) on which to build a superset conceptually similar to the d??R??S????d superset.

If sufficient prior knowledge is available, the anchor residues for test sets

30   can be chosen based on that prior knowledge. The choice of anchor positions and anchor residue identities for the RxxSF PKC-theta supersets (FIG. 2 and FIG. 6) were based on prior knowledge of the inventor on PKC specificity in which the dominant residues that determine PKC specificity were believed to include arginine at P-3, arginine at P-2, phenylalanine at P+1, arginine at P+2

and arginine at P+3. Therefore, some or all of such previously identified residues and/or positions can be chosen for the anchor positions of a particular test set or superset of peptides.

5   The method of the invention also provides an approach referred to as "Optimal Residue Position Scanning" (ORPS) to experimentally determine good anchor residues when prior knowledge is insufficient. Details of ORPS are described in Example 9 and Example 12, and their use further illustrated in Example 14.

Choice of the query positions and amino acids at the query position.

10   In most embodiments each test set has only one query position. This assures that the difference between peptides in the test set can be clearly attributed to change in a single amino acid at a standardized position.

    Of importance in the current method is the fact that the query position does NOT need to be adjacent to either an anchor position or to a

15 phosphorylatable position. This contrasts with pervasive use by previous worker of query-like positions adjacent to anchor-like positions (and phosphorylatable-like positions) in methods using "systematic amino acid variation on template substrate" (SAaVoTS). Particularly notable is that the extensive work of Tegge and colleagues on finding optimal peptides/inhibitors was based on query

20 residues adjacent to fixed residues (for example Dostmann WR et al. 1999. Pharmacol Ther 82:373-387; Tegge W et al. 1995. Biochemistry 34:10569-10577; Tegge WJ et al. 1998. Methods Mol Biol 87:99-106). Thus, the current method incorporates new flexibility relative to the prior art of "systematic amino acid variation on template substrate" by placing a query position at any position

25 relative to the anchor and phosphorylatable positions.

    Any amino acid can be selected for placement at the query position. While in some embodiments all available amino acids are systematically placed and tested in the query position, in other embodiments only a subset of natural amino acids are selected for placement in the query position. Hence, in some

30 embodiments, the test set of peptides would include one peptide for each natural amino acid. In other embodiments, cysteine is eliminated and only nineteen alternative amino acid residues are used.

    In other embodiments, economy is achieved by assuming that amino acids can be subdivided into classes that are most similar in their functional

properties. For example, using this strategy, a "reduced set" of only about thirteen amino acid residues are alternatively placed in the query position, as illustrated by FIG. 2 and FIG. 6. For example, one of skill in the art may choose to eliminate glutamic acid (E) by virtue of its similarity to aspartic acid (D);

5   isoleucine (I), methionine (M) and valine (V) can be eliminated by virtue of their similarity to leucine (L) and tyrosine (Y) can be eliminated by virtue of similarity to phenylalanine (F) (see further details in Example 2).

Choosing Residues and Conditions for Degenerate Positions

The degenerate amino acid position in the peptide pools can be created

10   such that any one of the twenty amino acids can occupy that position. However, this strategy can be altered by one of skill in the art to suit the needs of a particular test or situation. For example, one of skill in the art may elect not to use cysteine because can give rise to disulfide bonds and dimer formation.

In other embodiments, residues that may be phosphorylated (e.g. S, T,

15   and Y) can be excluded from the degenerate positions. However, serine, threonine and tyrosine residues may also be included because they can have a role in determining substrate specificity and because an experimental design minimizes noise when such residues are used in degenerate position. For example, in the methods of the invention noise from degenerate position serine,

20   threonine or tyrosine residues is minimized because of the abundance of the selected serine, threonine, or tyrosine residue at the P0 position relative to the rarity of these amino acids in degenerate positions. Moreover, phosphorylation at the P0 position is selectively enhanced by the anchor residues that guide the kinase to phosphorylate the appropriate residue. Hence, the types and positions

25   of degenerate residues can be varied as needed.

Two approaches can be used for inserting a degenerate set of amino acids into selected positions of a peptide. In one embodiment, a mixture of selected amino acid residues is added by a specific coupling step to create a degenerate position. However, different amino acid residues have different coupling

30   efficiencies and therefore, if equal amounts of each amino acid are used, each amino acid residue may not be equivalently represented at the degenerate position. The different coupling efficiencies of different amino acids can be compensated for by using a "weighted" mixture of amino acids at a coupling step, wherein amino acids with lower coupling efficiencies are present in greater

abundance than amino acids with higher coupling efficiencies. Conditions of the coupling can also be varied to facilitate achievement of a desired mix in the synthesized peptide. For example relatively low molar ratios minimize skewing by different coupling efficiencies; also, repetitive additions of low molar ratios

5   can augment efficiency while minimizing skewing.

In an alternative embodiment, the resin upon which the peptides are synthesized is divided into equivalent portions and then each portion is subjected to a separate coupling reaction that employs a distinct type of amino acid. After this coupling reaction, the resin aliquots are recombined and the procedure is

10   repeated for each degenerate position. This approach results in approximately equivalent representation of each different amino acid residue at the degenerate position.

The abundance of residues at the degenerate positions in the peptides can be controlled by a variety of different strategies (see FIG. 14). One procedure

15   for controlling the abundance of residues at the degenerate position is shown as plan 1 in FIG. 14, where an equal abundance of each amino acid residue is selected for each position. However, in many embodiments the abundance of amino acids is based on prior knowledge of the abundance of residues in human proteins or relevant regions thereof. One such embodiment utilized the average

20   abundance of various amino acids in the human proteome. The abundance of amino acids in human proteins was determined by reference to sequences tabulated by the National Center for Biotechnology Information (Plan 2, FIG. 14).

In another embodiment, the abundance of various amino acids at a

25   degenerate position correlates with the abundance of that amino acid in known kinase substrates (Plan 3, FIG. 14). Plan 3 of FIG. 14 takes into account the physiological relevance of various residues and resembles the residue abundance found in physiologic substrates for the kinase(s). To this end, the inventor has accumulated a list of known or suspected substrate sites for PKC and has

30   determined the residue frequency in the regions surrounding those sites (Plan 3, FIG. 14). The intent was to create a method that screens the most relevant peptide sequences for targeted biological processes.

Hence, in some embodiments a degenerate mixture of residues is used that is like the types of amino acid residues thought to be most relevant to a

particular kinase. Implementing this improvement by deviating from equal
abundance is not a problem in the present method but could be a problem in
prior art approaches (e.g. U.S. Patent 6,004,757 to Cantley) because prior art
approaches depend on detection of substrate residue by sequence analysis of the

5   phosphorylated product and a low abundance of a particular residue in the
degenerate peptide pool being phosphorylated would decrease the reliability of
detecting such a difference.

Additional residues beyond the core peptide

The peptide pools in a test set or in a superset can include additional

10   residues at either the N-terminus or C-terminus (or both). Such additional amino
acid residues may provide additional attachment points or other functions useful
to one of skill in the art. For example, in the ninety peptide test set having the
formula Rxx-S-F (FIG. 2), each peptide included a three residue N-terminal
linker of biotinylated lysine, dansylated lysine and glycine. The biotin moiety

15   provided an efficient mechanism for capture of the peptide before, during or
after an assay. The dansyl moiety also provided a convenient means to quantify
the amount of each peptide by measuring light absorption at 335 nm. The
glycine provided flexibility in connecting the linker to the remainder of the
peptide. Hence, such linkers can be used in the methods, articles and kits of the

20   invention.

Examples of other variations in tests sets of peptides

The number of peptide pools in a test set can vary. In some
embodiments, the number of peptide pools in the test set is equivalent to the
number of amino acids tested at the query position. Hence, for example, if all

25   twenty naturally-occurring amino acids are tested in the test set, the number of
peptide pools would be twenty. However, in many embodiments, fewer than
twenty amino acids are tested because one of skill in the art may have
information indicating that certain amino acids need not be tested. Moreover,
many amino acid analogs are available to one of skill in the art and in some

30   instances the skilled artisan may choose to test such an amino acid analog at the
query position. In such instances, amino acid analogs may be used in the test
sets of the invention and the number of peptide pools can be greater than twenty.
Also, under special circumstances it is useful to use a mixture of amino acids,
such as (R + K) or (D + E) instead of a single amino acid at a query position.

Similarly, special circumstances may dictate use of a limited mix of amino acids at the phosphorylatable position (such as S + T), or at an anchor position (such as I + L + M + V). Note that FIG. 2 illustrates that the same degenerate peptide can be used in three different sets: for example, the peptide symbolized by

5   'ddddRdd-S-Fdd' (shaded) was an element of the P-3 set, the P-0 set, and the P+1 set.

The number of test sets in a superset or collection of peptide pools can also vary. In general a superset has at least two test sets of peptide pools. Typically the number of test sets corresponds to the number of positions around

10   the phosphorylation site that are being tested, which is usually in the range of from about five to about twenty positions (or test sets). Moreover, a given test set can be used as part of different supersets. Also, practical considerations such as number of wells in a standardized plate (e.g. 96 or 384) often contribute to the choices made regarding number peptide pools in a test set, and number of test

15   sets in a superset. Moreover, different test sets can be used as part of different supersets.

The length of a peptide in a peptide pool can also vary. For example, although the amino acid sequences described in this application are often about five to about fifteen amino acids in length, a peptide that is shorter than five

20   amino acids may be used in some embodiments. For example, a peptide as short as about three amino acids in length may be used as a substrate. The upper size of the peptides used in the test sets and supersets is not critical and can vary as desired by one of skill in the art. However, peptides that are chemically synthesized become more expensive as their length increases. Hence, one of

25   skill in the art may choose to limit the size of the peptides employed to about 100 or fewer amino acids, or about 50 or fewer amino acids, or about 30 or fewer amino acids, or about 25 or fewer amino acids.

In some embodiments the peptide pools used in the test sets and supersets of the invention are soluble pools of peptides. The term "soluble peptide pools"

30   is intended to mean a population of peptides that are not attached to a solid support at the time they are subjected to phosphorylation.

In alternative embodiments, the peptides used in the test sets and supersets of the invention can be attached to a solid support such as a bead, a well of a microtiter dish, a membrane or a plastic pin. For general descriptions of

the construction of solid-support bound peptide libraries see for example

Geysen, H. M., et al. (1986) Mol. Immunol. 23:709-715; Lam, K. S., et al.

(1991) Nature 354:82-84; and Pinilla, C., et al. (1992) BioTechniques 13:901-

905. For this type of library, the peptides can be synthesized while attached to a

5   solid support such as a bead, and degenerate positions are created by splitting the

population of beads, coupling different amino acids to different subpopulations

and recombining the beads. The final product is a population of beads each

carrying many copies of a single unique peptide. This approach has been termed

"one bead/one peptide".

10   The choice of a soluble versus immobilized format should not be based

solely on convenience of the assay; some studies conducted by the inventors

suggest that significant differences in specificity are observed with the same

peptides assayed in solution versus assays performed on immobilized peptides.

Therefore, the distinction between soluble and immobilized may be of

15   considerable importance. The use of soluble peptide pools as the preferred

embodiment of this invention distinguishes the invention from many prior

methods performed with immobilized peptides. Also, those of skill in the art

should carefully assess all the implications of these alternative formats when

choosing the design of test sets of peptides for particular applications.

20   The peptides utilized in the test sets and supersets of the invention can be

prepared by any method available to one of skill in the art. For example, the

peptides can be constructed by *in vitro* chemical synthesis, for example using an

automated peptide synthesizer. As described herein the peptides can be soluble

peptide pools or the peptides can be attached to a solid support such as a bead,

25   membrane, microtiter well, tube or other convenient solid support.

Standard techniques for *in vitro* chemical synthesis of peptides are

known in the art. For example, peptides can be synthesized by

(benzotriazolyloxy)tris (dimethylamino)-phosphonium hexafluorophosophate

(BOP)/1-hydroxybenzotriazole coupling protocols. Automated peptide

30   synthesizers are commercially available (e.g., Milligen /Biosearch 9600). For

general descriptions of the construction of soluble synthetic peptide libraries see

for example Houghten, R. A., et al., (1991) Nature 354:84-86 and Houghten, R.

A., et al., (1992) BioTechniques 13:412-421.

Analysis of kinase specificity with non-degenerate peptides

Although degenerate peptides are particularly useful for studying kinase peptide specificity, strategic use of non-degenerate peptides can also be effective for identifying new substrates (Tables 3, 4, 5, 9). The present invention also teaches strategic design of sets of single sequence peptides (i.e. no degenerate

5 positions) so that they can be used for elucidating kinase peptide specificity of basophilic kinases (Example 13 and Example 14).

Binding Entities that Bind to Substrates of Kinases

The invention also contemplates binding entities that can bind to peptides or proteins that may be phosphorylated by a kinase. In some embodiments, the

10 binding entities bind to the non-phosphorylated substrate; in other embodiments the binding entities bind to phosphorylated substrates.

For example, as illustrated herein, a site-specific phospho-antibody was generated and used to detect phosphorylation at a specific peptidyl sequence. A phospho-peptide having sequence CDKEKSKG-(pS)-LKRK-OH (SEQ ID

15 NO:570) was made. This sequence (without phosphorylation) comprises the C-terminus of SHP-1 and was chosen for study because the methods of the current invention predicted that it was a candidate site for phosphorylation by PKC (see Example 10). This phospho-peptide includes a sequence that corresponds to the C-terminus of SHP-1 but, in addition, it has an N-terminal cysteine useful for

20 coupling to a carrier. The corresponding non-phosphorylated peptide was also synthesized for use as a control. The phospho-peptide (SEQ ID NO:570) was coupled onto a KLH carrier, rabbits were immunized, and anti-sera samples were screened for reactivity with the SEQ ID NO:570 phospho-peptide by ELISA assay. Antibodies reactive with corresponding non-phosphorylated peptide were

25 removed from anti-sera by passing the anti-sera through a column having the non-phosphorylated peptide bound to the column matrix. Finally, anti-sera were enriched for phospho-specific reactivity by use of an affinity column made from the phospho-peptide. The antibody preparation so produced was called the anti-pS591 antibody preparation.

30 The specificity of the antibody for SHP-1 pS591 was confirmed by Western blot analysis (see FIG. 43). When the anti-SHP-1 pS591 antibody was used at a dilution of 1:15,000, only a single strong band was detected on a Western blot of a lysate of Jurkat cells. The position of this band was characteristic of SHP-1. In contrast, in similar experiments, an antibody that

binds generally to sites phosphorylated by PKC bound to many bands. This antibody facilitated studies of the functional importance of phosphorylation of this site in SHP-1 (see Example 10).

Thus the invention provides binding entities that can selectively bind to
5   sites that are phosphorylated by various kinases. In other embodiments, the binding entities selectively bind to non-phosphorylated sites that normally are recognized by kinases. Such binding entities can be used *in vitro* or *in vivo* for detecting phosphorylated or non-phosphorylated peptides or proteins or for modulating the function of a phosphorylated or non-phosphorylated protein. As
10   used herein, a binding entity is any small molecule, peptide, or polypeptide that can bind to a peptidyl substrate site of kinase. In some embodiments, the binding entities are antibodies.

Hence, binding entities can bind to a phosphorylated peptidyl substrate sequence but exhibit significantly less or substantially no binding to the
15   corresponding non-phosphorylated peptidyl substrate sequence. Binding entities of the invention can also bind to a non-phosphorylated peptidyl substrate sequence but exhibit significantly less or substantially no binding to the corresponding phosphorylated peptidyl substrate sequence.

For example, binding entities and antibodies contemplated by the
20   invention may bind to a peptide having a combination of SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517 or 570. In another embodiment, binding entities and antibodies of the invention bind to a peptide having SEQ ID NO:76,
25   81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517, or 570, but not any other of the peptides. In further embodiments of the invention, binding entities and antibodies of the invention bind to a phosphorylated peptide having one of SEQ
30   ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517 or 570, but exhibit significantly less or substantially no binding to the corresponding non-phosphorylated peptidyl substrate sequence.

In still further embodiments of the invention, binding entities and
antibodies of the invention bind to a non-phosphorylated peptide having one of
SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113,
121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171,

5   173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517 or 570, but exhibit
significantly less or substantially no binding to the corresponding
phosphorylated peptidyl substrate sequence.

In some embodiments, the binding entities recognize phosphorylated or
non-phosphorylated peptidyl sequences having any one of SEQ ID NO: 89, 102,

10   110, 112, 127, 177, 182, 209, 474-488 or 489. In other embodiments, the
binding entities recognize phosphorylated or non-phosphorylated peptidyl
sequences having any one of SEQ ID NO: 173, 185, 192, 196, 200, 490-491 or
492.

In further embodiments, the binding entities further differentiate between

15   a phosphorylated peptide having any one of SEQ ID NO: 298, 301-324, 326-347,
349-400, 402-410, 412-473, 571-643 or 644, and a non-phosphorylated peptide
that differs from the phosphorylated peptide by substitution of Ser for the pSer
or substitution of a Thr for the pThr. For example, such a phosphorylated
peptide can have any one of SEQ ID: 298, 320, 324, 350, 351, 366, 388, 394,

20   398, 402, 418, 464, 571-595 or 596. In other embodiments, the phosphorylated
peptide can have any one of SEQ ID: 301, 310, 317, 322, 344, 352, 371, 406,
597-599 or 600. One example of a preferred binding entity of the invention is a
binding entity that binds to a phosphorylated peptide that includes SEQ ID
NO:298. Another example of a preferred binding entity of the invention is a

25   binding entity that binds to a phosphorylated peptide that includes SEQ ID
NO:313 or 314. Another example of a preferred binding entity of the invention
is a binding entity that binds to a phosphorylated peptide that includes SEQ ID
NO:361 or 362.

The invention provides antibodies and binding entities made by available

30   procedures that can bind a non-phosphorylated peptide or phosphorylated
peptide of the invention. The binding domains of such antibodies, for example,
the CDR regions of these antibodies, can also be transferred into or utilized with
any convenient binding entity backbone.

Antibody molecules belong to a family of plasma proteins called immunoglobulins, whose basic building block, the immunoglobulin fold or domain, is used in various forms in many molecules of the immune system and other biological recognition systems. A standard antibody is a tetrameric

5   structure consisting of two identical immunoglobulin heavy chains and two identical light chains and has a molecular weight of about 150,000 daltons.

The heavy and light chains of an antibody consist of different domains. Each light chain has one variable domain (VL) and one constant domain (CL), while each heavy chain has one variable domain (VH) and three or four constant

10   domains (CH). See, e.g., Alzari, P. N., Lascombe, M.-B. & Poljak, R. J. (1988) *Three-dimensional structure of antibodies*. Annu. Rev. Immunol. 6, 555-580. Each domain, consisting of about 110 amino acid residues, is folded into a characteristic β-sandwich structure formed from two β-sheets packed against each other, the immunoglobulin fold. The VH and VL domains each have three

15   complementarity determining regions (CDR1-3) that are loops, or turns, connecting β-strands at one end of the domains. The variable regions of both the light and heavy chains generally contribute to antigen specificity, although the contribution of the individual chains to specificity is not always equal. Antibody molecules have evolved to bind to a large number of molecules by using six

20   randomized loops (CDRs).

Immunoglobulins can be assigned to different classes depending on the amino acid sequences of the constant domain of their heavy chains. There are at least five (5) major classes of immunoglobulins: IgA, IgD, IgE, IgG and IgM. Several of these may be further divided into subclasses (isotypes), for example,

25   IgG-1, IgG-2, IgG-3 and IgG-4; IgA-1 and IgA-2. The heavy chain constant domains that correspond to the IgA, IgD, IgE, IgG and IgM classes of immunoglobulins are called alpha (α), delta (δ), epsilon (ε), gamma (γ) and mu (μ), respectively. The light chains of antibodies can be assigned to one of two clearly distinct types, called kappa (κ) and lambda (λ), based on the amino

30   sequences of their constant domain. The subunit structures and three-dimensional configurations of different classes of immunoglobulins are well known.

The term "variable" in the context of variable domain of antibodies, refers to the fact that certain portions of variable domains differ extensively in

sequence from one antibody to the next. The variable domains are for binding
and determine the specificity of each particular antibody for its particular
antigen. However, the variability is not evenly distributed through the variable
domains of antibodies. Instead, the variability is concentrated in three segments

5 . called complementarity determining regions (CDRs), also known as
hypervariable regions in both the light chain and the heavy chain variable
domains.

The more highly conserved portions of variable domains are called
framework (FR) regions. The variable domains of native heavy and light chains

10 each comprise four FR regions, largely adopting a β-sheet configuration,
connected by three CDRs, which form loops connecting, and in some cases
forming part of, the β-sheet structure. The CDRs in each chain are held together
in close proximity by the FR regions and, with the CDRs from another chain,
contribute to the formation of the antigen-binding site of antibodies. The

15 constant domains are not involved directly in binding an antibody to an antigen,
but exhibit various effector functions, such as participation of the antibody in
antibody-dependent cellular toxicity.

An antibody that is contemplated for use in the present invention thus can
be in any of a variety of forms, including a whole immunoglobulin, an antibody

20 fragment such as Fv, Fab, and similar fragments, a single chain antibody which
includes the variable domain complementarity determining regions (CDR), and
the like forms, all of which fall under the broad term "antibody", as used herein.
The present invention contemplates the use of any specificity of an antibody,
polyclonal or monoclonal, and is not limited to antibodies that recognize and

25 immunoreact with a specific peptide sequence described herein or a derivative
thereof.

Moreover, the binding regions, or CDR, of antibodies can be placed
within the backbone of any convenient binding entity polypeptide. In preferred
embodiments, in the context of methods described herein, an antibody, binding

30 entity or fragment thereof is used that is immunospecific for any of the peptides
described herein, as well as the derivatives thereof, including the phosphorylated
derivatives thereof.

The term "antibody fragment" refers to a portion of a full-length
antibody, generally the antigen binding or variable region. Examples of antibody

fragments include Fab, Fab', F(ab')$_2$ and Fv fragments. Papain digestion of antibodies produces two identical antigen binding fragments, called Fab fragments, each with a single antigen binding site, and a residual Fc fragment. Fab fragments thus have an intact light chain and a portion of one heavy chain.

5   Pepsin treatment yields an F(ab')$_2$ fragment that has two antigen binding fragments that are capable of cross-linking antigen, and a residual fragment that is termed a pFc' fragment. Fab' fragments are obtained after reduction of a pepsin digested antibody, and consist of an intact light chain and a portion of the heavy chain. Two Fab' fragments are obtained per antibody molecule. Fab'

10   fragments differ from Fab fragments by the addition of a few residues at the carboxyl terminus of the heavy chain CH1 domain including one or more cysteines from the antibody hinge region.

Fv is the minimum antibody fragment that contains a complete antigen recognition and binding site. This region consists of a dimer of one heavy and

15   one light chain variable domain in a tight, non-covalent association (V$_H$ -V$_L$ dimer). It is in this configuration that the three CDRs of each variable domain interact to define an antigen binding site on the surface of the V$_H$ -V$_L$ dimer. Collectively, the six CDRs confer antigen binding specificity to the antibody. However, even a single variable domain (or half of an Fv comprising only three

20   CDRs specific for an antigen) has the ability to recognize and bind antigen, although at a lower affinity than the entire binding site. As used herein, "functional fragment" with respect to antibodies, refers to Fv, F(ab) and F(ab')$_2$ fragments.

Additional fragments can include diabodies, linear antibodies, single-

25   chain antibody molecules, and multispecific antibodies formed from antibody fragments. Single chain antibodies are genetically engineered molecules containing the variable region of the light chain, the variable region of the heavy chain, linked by a suitable polypeptide linker as a genetically fused single chain molecule. Such single chain antibodies are also referred to as "single-chain Fv"

30   or "sFv" antibody fragments. Generally, the Fv polypeptide further comprises a polypeptide linker between the VH and VL domains that enables the sFv to form the desired structure for antigen binding. For a review of sFv see Pluckthun in The Pharmacology of Monoclonal Antibodies, vol. 113, Rosenburg and Moore eds. Springer-Verlag, N.Y., pp. 269-315 (1994).

The term "diabodies" refers to a small antibody fragments with two antigen-binding sites, where the fragments comprise a heavy chain variable domain (VH) connected to a light chain variable domain (VL) in the same polypeptide chain (VH-VL). By using a linker that is too short to allow pairing

5   between the two domains on the same chain, the domains are forced to pair with the complementary domains of another chain and create two antigen-binding sites. Diabodies are described more fully in, for example, EP 404,097; WO 93/11161, and Hollinger et al., Proc. Natl. Acad Sci. USA 90: 6444-6448 (1993).

Antibody fragments contemplated by the invention are therefore not full-

10   length antibodies. However, such antibody fragments can have similar or improved immunological properties relative to a full-length antibody. Such antibody fragments may be as small as about 4 amino acids, 5 amino acids, 6 amino acids, 7 amino acids, 9 amino acids, about 12 amino acids, about 15 amino acids, about 17 amino acids, about 18 amino acids, about 20 amino acids,

15   about 25 amino acids, about 30 amino acids or more.

In general, an antibody fragment of the invention can have any upper size limit so long as it is has similar or improved immunological properties relative to an antibody that binds with specificity to a peptide or phosphorylated peptide described herein. For example, smaller binding entities and light chain antibody

20   fragments can have less than about 200 amino acids, less than about 175 amino acids, less than about 150 amino acids, or less than about 120 amino acids if the antibody fragment is related to a light chain antibody subunit. Moreover, larger binding entities and heavy chain antibody fragments can have less than about 425 amino acids, less than about 400 amino acids, less than about 375 amino

25   acids, less than about 350 amino acids, less than about 325 amino acids or less than about 300 amino acids if the antibody fragment is related to a heavy chain antibody subunit.

Antibodies directed against disease markers can be made by any available procedure. Methods for the preparation of polyclonal antibodies are

30   available to those skilled in the art. See, for example, Green, et al., Production of Polyclonal Antisera, in: Immunochemical Protocols (Manson, ed.), pages 1-5 (Humana Press); Coligan, et al., Production of Polyclonal Antisera in Rabbits, Rats Mice and Hamsters, in: Current Protocols in Immunology, section 2.4.1 (1992), which are hereby incorporated by reference.

Monoclonal antibodies can also be employed in the invention. The term "monoclonal antibody" as used herein refers to an antibody obtained from a population of substantially homogeneous antibodies. In other words, the individual antibodies comprising the population are identical except for

5   occasional naturally occurring mutations in some antibodies that may be present in minor amounts. Monoclonal antibodies are highly specific, being directed against a single antigenic site. Furthermore, in contrast to polyclonal antibody preparations that typically include different antibodies directed against different determinants (epitopes), each monoclonal antibody is directed against a single

10   determinant on the antigen. In additional to their specificity, the monoclonal antibodies are advantageous in that they are synthesized by the hybridoma culture, uncontaminated by other immunoglobulins. The modifier "monoclonal" indicates the character of the antibody indicates the character of the antibody as being obtained from a substantially homogeneous population of antibodies, and

15   is not to be construed as requiring production of the antibody by any particular method.

The monoclonal antibodies herein specifically include "chimeric" antibodies in which a portion of the heavy and/or light chain is identical or homologous to corresponding sequences in antibodies derived from a particular

20   species or belonging to a particular antibody class or subclass, while the remainder of the chain(s) is identical or homologous to corresponding sequences in antibodies derived from another species or belonging to another antibody class or subclass. Fragments of such antibodies can also be used, so long as they exhibit the desired biological activity. See U.S. Patent No. 4,816,567; Morrison

25   et al. Proc. Natl. Acad Sci. 81, 6851-55 (1984). The monoclonal antibodies herein also specifically include those made from different animal species, including mouse, rat, human and rabbit.

The preparation of monoclonal antibodies likewise is conventional. See, for example, Kohler & Milstein, Nature, 256:495 (1975); Coligan, et al., sections

30   2.5.1-2.6.7; and Harlow, et al., in: Antibodies: A Laboratory Manual, page 726 (Cold Spring Harbor Pub. (1988)), which are hereby incorporated by reference. Monoclonal antibodies can be isolated and purified from hybridoma cultures by a variety of well-established techniques. Such isolation techniques include affinity chromatography with Protein-A Sepharose, size-exclusion

chromatography, and ion-exchange chromatography. See, e.g., Coligan, et al., sections 2.7.1-2.7.12 and sections 2.9.1-2.9.3; Barnes, et al., Purification of Immunoglobulin G (IgG), in: Methods in Molecular Biology, Vol. 10, pages 79-104 (Humana Press (1992).

5   Methods of *in vitro* and *in vivo* manipulation of antibodies are available to those skilled in the art. For example, the monoclonal antibodies to be used in accordance with the present invention may be made by the hybridoma method as described above or may be made by recombinant methods, e.g., as described in U.S. Pat. No. 4,816,567. Monoclonal antibodies for use with the present

10   invention may also be isolated from phage antibody libraries using the techniques described in Clackson et al. Nature 352: 624-628 (1991), as well as in Marks et al., J. Mol Biol. 222: 581-597 (1991).

Methods of making antibody fragments are also known in the art (see for example, Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring

15   Harbor Laboratory, New York, (1988), incorporated herein by reference). Antibody fragments of the present invention can be prepared by proteolytic hydrolysis of the antibody or by expression of nucleic acids encoding the antibody fragment in a suitable host. Antibody fragments can be obtained by pepsin or papain digestion of whole antibodies conventional methods. For

20   example, antibody fragments can be produced by enzymatic cleavage of antibodies with pepsin to provide a 5S fragment described as F(ab')$_2$. This fragment can be further cleaved using a thiol reducing agent, and optionally using a blocking group for the sulfhydryl groups resulting from cleavage of disulfide linkages, to produce 3.5S Fab' monovalent fragments. Alternatively,

25   enzymatic cleavage using pepsin produces two monovalent Fab' fragments and an Fc fragment directly. These methods are described, for example, in U.S. Patents No. 4,036,945 and No. 4,331,647, and references contained therein. These patents are hereby incorporated by reference in their entireties.

Other methods of cleaving antibodies, such as separation of heavy chains

30   to form monovalent light-heavy chain fragments, further cleavage of fragments, or other enzymatic, chemical, or genetic techniques may also be used, so long as the fragments bind to the antigen that is recognized by the intact antibody. For example, Fv fragments comprise an association of V_H and V_L chains. This association may be noncovalent or the variable chains can be linked by an

intermolecular disulfide bond or cross-linked by chemicals such as glutaraldehyde. Preferably, the Fv fragments comprise V_H and V_L chains connected by a peptide linker. These single-chain antigen binding proteins (sFv) are prepared by constructing a structural gene comprising DNA sequences

5   encoding the V_H and V_L domains connected by an oligonucleotide. The structural gene is inserted into an expression vector, which is subsequently introduced into a host cell such as *E. coli*. The recombinant host cells synthesize a single polypeptide chain with a linker peptide bridging the two V domains. Methods for producing sFvs are described, for example, by Whitlow, et al.,

10 Methods: a Companion to Methods in Enzymology, Vol. 2, page 97 (1991); Bird, et al., Science 242:423-426 (1988); Ladner, et al, US Patent No. 4,946,778; and Pack, et al., Bio/Technology 11:1271-77 (1993).

Another form of an antibody fragment is a peptide coding for a single complementarity-determining region (CDR). CDR peptides ("minimal

15 recognition units") are often involved in antigen recognition and binding. CDR peptides can be obtained by cloning or constructing genes encoding the CDR of an antibody of interest. Such genes are prepared, for example, by using the polymerase chain reaction to synthesize the variable region from RNA of antibody-producing cells. See, for example, Larrick, et al., Methods: a

20 Companion to Methods in Enzymology, Vol. 2, page 106 (1991).

The invention contemplates human and humanized forms of non-human (e.g. murine) antibodies. Such humanized antibodies are chimeric immunoglobulins, immunoglobulin chains or fragments thereof (such as Fv, Fab, Fab', F(ab')$_2$ or other antigen-binding subsequences of antibodies) that contain

25 minimal sequence derived from non-human immunoglobulin. For the most part, humanized antibodies are human immunoglobulins (recipient antibody) in which residues from a complementary determining region (CDR) of the recipient are replaced by residues from a CDR of a nonhuman species (donor antibody) such as mouse, rat or rabbit having the desired specificity, affinity and capacity.

30 In some instances, Fv framework residues of the human immunoglobulin are replaced by corresponding non-human residues. Furthermore, humanized antibodies may comprise residues that are found neither in the recipient antibody nor in the imported CDR or framework sequences. These modifications are made to further refine and optimize antibody performance. In general,

humanized antibodies will comprise substantially all of at least one, and typically two, variable domains, in which all or substantially all of the CDR regions correspond to those of a non-human immunoglobulin and all or substantially all of the FR regions are those of a human immunoglobulin

5   consensus sequence. The humanized antibody optimally also will comprise at least a portion of an immunoglobulin constant region (Fc), typically that of a human immunoglobulin. For further details, see: Jones et al., Nature 321, 522-525 (1986); Reichmann et al., Nature 332, 323-329 (1988); Presta, Curr. Op. Struct. Biol. 2, 593-596 (1992); Holmes, et al., J. Immunol., 158:2192-2201

10   (1997) and Vaswani, et al., Annals Allergy, Asthma & Immunol., 81:105-115 (1998).

While standardized procedures are available to generate antibodies, the size of antibodies, the multi-stranded structure of antibodies and the complexity of six binding loops present in antibodies constitute a hurdle to the improvement

15   and the manufacture of large quantities of antibodies. Hence, the invention further contemplates using binding entities, which comprise polypeptides that can recognize and bind to kinase substrates provided herein.

A number of proteins can serve as protein scaffolds to which binding domains can be attached and thereby form a suitable binding entity. The binding

20   domains bind or interact with the peptide sequences of the invention while the protein scaffold merely holds and stabilizes the binding domains so that they can bind. A number of protein scaffolds can be used. For example, phage capsid proteins can be used. See Review in Clackson & Wells, Trends Biotechnol. 12:173-184 (1994). Phage capsid proteins have been used as scaffolds for

25   displaying random peptide sequences, including bovine pancreatic trypsin inhibitor (Roberts et al., PNAS 89:2429-2433 (1992)), human growth hormone (Lowman et al., Biochemistry 30:10832-10838 (1991)), Venturini et al., Protein Peptide Letters 1:70-75 (1994)), and the IgG binding domain of Streptococcus (O'Neil et al., Techniques in Protein Chemistry V (Crabb, L,. ed.) pp. 517-524,

30   Academic Press, San Diego (1994)). These scaffolds have displayed a single randomized loop or region that can be modified to include binding domains for kinase substrates.

Researchers have also used the small 74 amino acid α-amylase inhibitor Tendamistat as a presentation scaffold on the filamentous phage M13. McConnell, S.

J., & Hoess, R. H., J.Mol. Biol. 250:460-470 (1995). Tendamistat is a β-sheet protein from *Streptomyces tendae*. It has a number of features that make it an attractive scaffold for binding entities, including its small size, stability, and the availability of high resolution NMR and X-ray structural data. The overall topology of Tendamistat is

5   similar to that of an immunoglobulin domain, with two β-sheets connected by a series of loops. In contrast to immunoglobulin domains, the β-sheets of Tendamistat are held together with two rather than one disulfide bond, accounting for the considerable stability of the protein. The loops of Tendamistat can serve a similar function to the CDR loops found in immunoglobulins and can be easily randomized by *in vitro*

10   mutagenesis. Tendamistat is derived from Streptomyces tendae and may be antigenic in humans. Hence, binding entities that employ Tendamistat are preferably employed *in vitro*.

Fibronectin type III domain has also been used as a protein scaffold to which binding entities can be attached. Fibronectin type III is part of a large subfamily (Fn3

15   family or s-type Ig family) of the immunoglobulin superfamily. Sequences, vectors and cloning procedures for using such a fibronectin type III domain as a protein scaffold for binding entities (e.g. CDR peptides) are provided, for example, in U.S. Patent Application Publication 20020019517. See also, Bork, P. & Doolittle, R. F. (1992) Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad.

20   Sci. USA 89, 8990-8994; Jones, E. Y. (1993) The immunoglobulin superfamily Curr. Opinion Struct. Biol. 3, 846-852; Bork, P., Hom, L. & Sander, C. (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. J. Mol. Biol. 242, 309-320; Campbell, I. D. & Spitzfaden, C. (1994) Building proteins with fibronectin type III modules Structure 2, 233-337; Harpez, Y. & Chothia, C.

25   (1994).

In the immune system, specific antibodies are selected and amplified from a large library (affinity maturation). The combinatorial techniques employed in immune cells can be mimicked by mutagenesis and generation of combinatorial libraries of binding entities. Variant binding entities, antibody fragments and antibodies therefore

30   can also be generated through display-type technologies. Such display-type technologies include, for example, phage display, retroviral display, ribosomal display, and other techniques. Techniques available in the art can be used for generating libraries of binding entities, for screening those libraries and the selected binding entities can be subjected to additional maturation, such as affinity maturation. Wright

and Harris, supra., Hanes and Plucthau PNAS USA 94:4937-4942 (1997) (ribosomal

display), Parmley and Smith Gene 73:305-318 (1988) (phage display), Scott TIBS

17:241-245 (1992), Cwirla et al. PNAS USA 87:6378-6382 (1990), Russel et al. Nucl.

Acids Research 21:1081-1085 (1993), Hoganboom et al. Immunol. Reviews 130:43-68

5   (1992), Chiswell and McCafferty TIBTECH 10:80-84 (1992), and U.S. Pat. No.

5,733,743.

The invention therefore also provides methods of mutating antibodies, CDRs or

binding domains to optimize their affinity, selectivity, binding strength and/or other

desirable properties. A mutant binding domain refers to an amino acid sequence

10   variant of a selected binding domain (e.g. a CDR). In general, one or more of the

amino acid residues in the mutant binding domain is different from what is present in

the reference binding domain. Such mutant antibodies necessarily have less than 100%

sequence identity or similarity with the reference amino acid sequence. In general,

mutant binding domains have at least 75% amino acid sequence identity or similarity

15   with the amino acid sequence of the reference binding domain. Preferably, mutant

binding domains have at least 80%, more preferably at least 85%, even more preferably

at least 90%, and most preferably at least 95% amino acid sequence identity or

similarity with the amino acid sequence of the reference binding domain.

For example, affinity maturation using phage display can be utilized as one

20   method for generating mutant binding domains. Affinity maturation using phage

display refers to a process described in Lowman et al., Biochemistry 30(45): 10832-

10838 (1991), see also Hawkins et al., J. Mol Biol. 254: 889-896 (1992). While not

strictly limited to the following description, this process can be described briefly as

involving mutation of several binding domains or antibody hypervariable regions at a

25   number of different sites with the goal of generating all possible amino acid

substitutions at each site. The binding domain mutants thus generated are displayed in a

monovalent fashion from filamentous phage particles as fusion proteins. Fusions are

generally made to the gene III product of M13. The phage expressing the various

mutants can be cycled through several rounds of selection for the trait of interest, e.g.

30   binding affinity or selectivity. The mutants of interest are isolated and sequenced.

Such methods are described in more detail in U.S. Patent 5,750,373, U.S. Patent

6,290,957 and Cunningham, B. C. et al., EMBO J. 13(11), 2508-2515 (1994).

Therefore, in one embodiment, the invention provides methods of manipulating

binding entity or antibody polypeptides or the nucleic acids encoding them to generate

binding entities, antibodies and antibody fragments with improved binding properties
that recognize kinase substrate sequences.

Such methods of mutating portions of an existing binding entity or antibody
involve fusing a nucleic acid encoding a polypeptide that encodes a binding domain for
5   a disease marker to a nucleic acid encoding a phage coat protein to generate a
recombinant nucleic acid encoding a fusion protein, mutating the recombinant nucleic
acid encoding the fusion protein to generate a mutant nucleic acid encoding a mutant
fusion protein, expressing the mutant fusion protein on the surface of a phage, and
selecting phage that bind to a kinase substrate.

10   Accordingly, the invention provides antibodies, antibody fragments, and
binding entity polypeptides that can recognize and bind to a kinase substrate (e.g., a
peptide sequence having any of the peptidyl sequences described herein). The
invention further provides methods of manipulating those antibodies, antibody
fragments, and binding entity polypeptides to optimize their binding properties or other
15   desirable properties (e.g., stability, size, ease of use).

Such phospho-antibody production is well known to practitioners of the
art; pertinent descriptions of such approaches include those described in
CURRENT PROTOCOLS IN CELL BIOLOGY, Chap. 16. ANTIBODIES AS CELL
BIOLOGICAL TOOLS, unit 16.6 Production of Antibodies That Recognize Specific
20   Tyrosine-Phosphorylated Peptides. In particular, methods available in the art
include, purification of binding entities that bind specificity to the
phosphorylated peptide; depletion of binding entities that cross-react on the non-
phosphorylated peptide and depletion of binding entities that cross-react on the a
distinct phosphopeptide.

25   **Kinases that can be used in the Methods of the Invention**

The methods of the invention can be used to identify the specificity of
any type of wild type or mutant kinase from any prokaryotic or eukaryotic
species. For example, the kinase can be a protein-serine/threonine specific
kinase (in which case a peptide library or set with a fixed non-degenerate serine
30   or threonine is used), a protein-tyrosine specific kinase (in which case a peptide
library or set with a fixed non-degenerate tyrosine is used) or a dual-specificity
kinase (in which case a peptide library or set with either a fixed non-degenerate
serine, threonine or tyrosine can be used). Examples of protein kinases that can

be utilized in the methods of the invention can also be found in Hanks et al. (1988) Science 241:42-52 and Manning G et al. 2002. Science 298:1912-1934.

Protein-serine/threonine specific kinases that can be used in the methods of the invention include and of those listed herein as well as: 1) cyclic

5   nucleotide-dependent kinases, such as cyclic-AMP-dependent protein kinases (e.g., protein kinase A) and cyclic-GMP-dependent protein kinases; 2) calcium-phospholipid-dependent kinases, such as protein kinase C; 3) calcium-calmodulin-dependent kinases, including CaMII, phosphorylase kinase (PhK), myosin light chain kinases (e.g., MLCK-K, MLCK-M), PSK-H1 and PSK-C3;

10   4) the SNF1 family of protein kinases (e.g., SNF 1, nim1, KIN1 and KIN2); 5) casein kinases (e.g., CKII); 6) the Raf-Mos proto-oncogene family of kinases, including Raf, A-Raf, PKS and Mos; and 7) the STE7 family of kinases (e.g. STE7 and PBS2). Additionally, the protein-serine/threonine specific kinase can be a kinase involved in cell cycle control. Many kinases involved in cell cycle

15   control have been identified. Cell cycle control kinases include the cyclin dependent kinases, which are heterodimers of a cyclin and kinase (such as cyclin B/p33^{cdc2}, cyclin A/p33^{CDK2}, cyclin E/p33^{CDK2} and cyclin D1/p33^{CDK4}). Other cell cycle control kinases include Wee1 kinase, Nim1/Cdr1 kinase, Wis1 kinase and NIMA kinase.

20   Protein-tyrosine specific kinases that can be used in the methods of the invention include: 1) members of the src family of kinases, including pp60^{c-src}, pp60^{v-src}, Yes, Fgr, FYN, LYN, LCK, HCK, Dsrc64 and Dsrc28; 2) members of the Abl family of kinases, including Abl, ARG, Dash, Nabl and Fes/Fps; 3) members of the epidermal growth factor receptor (EGFR) family of kinases,

25   including EGFR, v-Erb-B, NEU and DER; 4) members of the insulin receptor (INS.R) family of growth factors, including INS.R, IGF1R, DILR, Ros, 7less, TRK and MET; 5) members of the platelet-derived growth factor receptor (PDGFR) family of kinases, including PDGFR, CSF1R, Kit and RET.

Other protein kinases which can be used in the method of the invention

30   include syk, ZAP70, Focal Adhesion Kinase, erk1, erk2, erk3, MEK, CSK, BTK, ITK, TEC, TEC-2, JAK-1, JAK-2, LET23, c-fms, S6 kinases (including p70^{S6} and RSKs), TGF-β/activin receptor family kinases and Clk.

Kits

The invention is further directed to a kit having a test set or an array of peptide pools for identifying kinase substrate specificities. The peptides used in the test sets and arrays can be soluble peptides or peptides attached to a solid

5 support. Instructions for using the array can also be included in the kit.

As described above, a test set contains peptide pools, wherein every peptide in each of the peptide pools has an amino acid that can be phosphorylated by a kinase, a query amino acid, at least one anchor amino acid, and at least one degenerate amino acid. The amino acid that can be

10 phosphorylated by a kinase is at a defined phosphorylation position and every peptide of every peptide pool within a test set of peptide pools has an identical amino acid that can be phosphorylated by a kinase in that phosphorylation position. The query amino acid is at a defined query position within a test set but the query amino acid's identity at that defined query position is

15 systematically varied from one peptide pool to the next peptide pool within a test set of peptide pools. Each anchor amino acid is at a defined anchor position within a test set and an identical anchor amino acid is present at that defined position in every peptide of every peptide pool in the test set, but each test set of the series of test sets can have different anchor amino acids. The at least one

20 degenerate amino acid is an unknown amino acid selected from a degenerate mixture of amino acids.

The methods and kits of the invention can be used to determine an amino acid sequence motif for the phosphorylation site of any kinase. The preferred embodiment of such kits includes software to facilitate calculation of results,

25 determination of derived parameters such as residue preference and scores for a position specific scoring matrix, and display of results in informative formats such as the PSSM Logo. The kits of the invention can also include any item, reagent or solution useful for performing the methods of the invention. Such items can include microtiter plates, arrays of peptide pools where the peptides

30 are attached to a solid support, tubes for diluting reagents, and the like. Reagents useful for performing the methods of the invention include, for example, ATP, γ-labeled ATP, cations and co-factors typically utilized by kinases. Solutions useful for performing the method include buffer solutions for controlling or

adjusting the pH of the kinase assay mixture, sterile deionized water for diluting and reconstituting reagents, and the like.

The invention is further illustrated by the following non-limiting Examples.

5

EXAMPLE 1: Peptide synthesis and *in vitro* kinase assay

Materials

DIEA, piperidine (peptide synthesis grade), and TFA (HPLC grade) were obtained from Chem-Impex (Wood Dale, IL). DMF, ACN, MTBE, and MeOH

10   were obtained from EM Science (Gibbstown, NJ). HOBT and HBTU (peptide synthesis grade) were obtained from AnaSpec (San Jose, CA). Fmoc-amino acid derivatives were obtained from AnaSpec (San Jose, CA) and Chem-Impex (Wood Dale, IL). Biotin was obtained from SynPep (Dublin, CA).

Peptide Synthesis

15   Peptides were synthesized as C-terminal amides on Mimotopes (Clayton, Australia) SynPhase Rink amide acrylic-grafted polypropylene solid support (loading 7.5 μmole), arranged in a 12 x 8 format, in 96 well microtiter plates. Amino acid solution delivery was facilitated by a PinPal Amino Acid Indexer to indicate the appropriate amino acid to be delivered for each peptide in each

20   coupling cycle. A solution containing a mixture of nineteen amino acids was delivered for specific peptides and coupling cycles to create degenerate peptides. Activation was preformed *in situ* with a solution of 0.1 M HOBT/HBTU/DIEA in DMF. Each unique peptide sequence was synthesized with an N terminal Biotin-Lys-Gly spacer. A dansyl group was attached to the side chain of the

25   spacer Lysine to serve as a chromophore (330 nm) to facilitate peptide quantification. Deprotection with 25% piperidine, DMF and methanol washes were preformed batch wise. After completion of the synthesis, the peptides were cleaved from the solid support and deprotected by acidolysis in the presence of scavengers using TFA/EDT/TA/anisole 90:4:3:3 (v/v/v/v). The crude peptides

30   were precipitated and washed three times with cold MTBE, and lyophilized from water/ACN/HOAc 8:1:1 (v/v/v).

Analysis

The peptide products were validated and quantified via high throughput LC-MS. The system consisted of a Shimadzu (Columbia, MD) VP series HPLC

system and a PE Sciex (Foster City, CA) API 165 single quadrapole mass spectrometer. Reverse phase separations of 1 μL injections were preformed using two Phenominex (Torrance, CA) 30 x 1.0 mm Luna 3 μ C8 columns at 50° C with a flow rate of 350 μL/min. The peptides were eluted by a linear gradient

5   from 0% to 60% MeOH (0.1% HOAc) over five minutes and detected at 330 nm and 220 nm. For each LCMS injection, (M+H)/Z was extracted from MS data and compared to the expected mass for that sample, as calculated from its sequence. The UV absorbance trace was integrated to determine purity and yield.

10   **Degenerate Peptide Quantification**

Absorbance data for 10 μL aliquots of degenerate peptide solution were acquired using a Labsystems (Beverly, MA) Multiskan Ascent plate reader equipped with a 340 nm filter. Yield was determined using a concentration factor calculated from absorbance data acquired on the same system from

15   samples of known concentration that also contained a dansyl chromophore.

Dried degenerate peptides were reconstituted in 90% water/10% ethanol. The concentration of peptide was determined by measurement of absorption at 335 nm (maximal absorption wavelength for dansyl group), stock diluted to 1mM and stored in sealed well at 4 °C. A replica plate was prepared with

20   peptides at 100μM concentration in 90% water/10% ethanol and stored similarly.

Kinase preparations

Catalytically active preparations of the kinases of interest were either purchased or prepared. Purchased and tested active kinase preparations including the following: PKC-alpha, PKC-delta, PKC-epsilon, PKC-zeta, PKC-

25   mu, PKA, PKG from Calbiochem, ROK alpha/ROCK-II, active from Upstate Biotechnology, and AKT1 from Panvera.

An example of the purification procedure used for production of active kinase is as follows. A preparation of PKC-theta was prepared using a Gateway expression construct containing PKC-theta that was expressed in baculovirus,

30   which were used to infect Sf9 cells. The cell pellet from a liter of baculovirus-infected Sf9 cells was resuspended in 20 volumes (60 ml) of extraction buffer (20 mM Na phosphate buffer pH 7.5, 500 mM NaCl, 5 mM pyrophosphate, 10% glycerol, 10 mM imidazole, 1 mM PMSF), sonicated twice for one minute (1 cm tip at 60% power and 50% duty cycle) and cell disruption was verified

microscopically. The sample was adjusted to five mM MgCl₂ and treated with

one unit benzonase/ml for an additional 20 minute on ice. The sample was

clarified by centrifugation in a JA-20 rotor at 15K for 30 min at 4 °C, filtered

through a 0.8 mm filter and applied at 0.5 ml/min to a one ml chelating

5   sepharose column previously charged with nickel and equilibrated with

extraction buffer. The column was washed with extraction buffer at one ml/min

to baseline and eluted in a 20 ml gradient (20-500 mM imidazole in extraction

buffer) into one ml fractions that were analyzed by SDS-PAGE. Fractions with

the highest concentration of protein were pooled, were dialyzed twice against

10   one liter of 20 mM Na PO4 pH 7.5, 50 mM NaCl buffer. The kinase pool was

dialyzed twice against 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM EDTA, 5

mM DTT, 0.05% Triton-X-100. After dialysis, the sample was adjusted to 50%

glycerol and quick-frozen in a dry ice/ethanol bath.

More than 20 other preparations of PKC-theta have also been prepared

15   and tested in the inventor's laboratory. The have been typically been transiently

expressed in HEK293 cells, and purified by His-tag based isolation conceptually

similar to that described above. Alternatively, they were immunoaffinity

purified using anti-HA tag antibody to capture the protein when it has been fused

to a HA epitope tag; such preps are released by incubation in an excess

20   concentration of HA peptide. These include preparations derived from more than

10 different variant constructs of PKC-theta. Point mutations have been

produced using the QuikChange system from Stratagene, using the

manufacturer's suggested procedures.

Kinase assay

25   The conditions of the kinase assay and the amount of active kinase used

varied with the kinase and with the accuracy needed. For a typical experiment,

5-20 ng of kinase was used per well and each peptide pool was assayed in

duplicate wells. Note that the absolute amount of kinase used was not usually a

critical parameter, because the desired information related to specificity of the

30   kinase not its absolute activity, and robustness of the assay depends on

comparisons of the same amount of kinase on different peptides. The

combination of kinase concentration and assay duration was modified to assure

that the stoichiometry of peptide phosphorylation never exceeded 5%. The

choice of kinase buffer depended on the kinase being analyzed. For studies of

PKC, 100mM HEPES, 0.05% Triton-X100, 1mM CaCl2, 20mM MgCl2,

0.2mg/ml phosphatidyl serine (Avanti Polar Lipids), PMA 100ng/ml was

typically used. The lipid stock was prepared by transferring 3mg phosphatidyl

serine into iced mixture of 450μl water plus 50μl of 10% Triton-X100,

5   sonicating 10 times on ice for 1 sec each.

        The kinase reaction mixture was assembled by sequential addition to a

tube held on ice of: 5μl peptide (100μM for final concentration of 10μM), 15μl

of kinase (typically 5ng/well, in appropriate kinase buffer), 30μl of ATP

(1uCi/well of ^{32}P-gamma ATP in a stock of 167 μM cold ATP in the kinase

10   buffer; for final concentration for 100μM ATP). The mixture was rapidly

warmed to desired reaction temperature (30°C for PKC) and incubated for the

desired duration (usually 10 minutes). The kinase assay was terminated by

transfer to 4°C water batch, and rapid addition of an equal volume (50μl) of stop

solution [0.1M ATP + 0.1M EDTA in water, pH 8].

15   The peptides were then captured from the reaction mixture by transfer to

a Reacti-Bind Streptavidin High Binding Capacity Coated Plates (HBC) (Pierce

Biotechnology) as follows. The HBC plates were pre-rinsed three times with

PBS/Tween PBS/Tween20 0.05% (PBS/Tween). Part of all of the reaction

mixture was then transferred wells of a HBC plate pre-filled with 90μl of

20   phosphate-buffered saline (PBS); typically each aliquots of each phosphorylation

reaction were transferred to duplicate HBC plates to assure accuracy by

additional replication

        For kinase assays done at the standard peptide concentration of 10μM,

the peptide concentration in the reaction mixture becomes 5μM after addition of

25   the stop solution; consequently 10μl of the reaction (50 pMoles of peptide) was

transferred to the HBC plate. More generally, the amount of reaction mixture

transferred was estimated to be about 50 pMoles of peptide. The inventor had

validated that 50 pMoles of peptide was reliably and completely captured by the

wells that had a nominal binding capacity of 125 pMoles. The HBC plates were

30   incubated for 0.5 to 1.5hr at room temperature for complete binding of

biotinylated peptides to plate-bound streptavidin. The HBC plates were then

washed extensively with PBS/Tween. Five washes were done routinely and

additional wash steps were added if the wash solution removed from the plate

had measurable radioactivity as detected using a Geiger counter. This step is

essential to obtaining a good the signal to noise ratio because the fraction of

radioactivity incorporated in the peptides was a tiny fraction of the total in the

reaction mixture. The wells were air-dried. A volume of 40 – 50 µl of

microScint-20 (Packard Instruments) was added to each well. The plates were

5   covered with stick-on film sheet. Radioactive emissions were measured in a

TopCount NXT Microplate Scintillation and Luminescence Counter (Packard

Instruments). Typically samples were counted for 5 minutes (or more) to

improve the signal to noise ratio when counts were low.

10   **EXAMPLE 2: Use of Reduced set of Query Residues**

The methods described herein provide for systematic variation of the

query amino acid between peptides pools of a test set. In one embodiment, all

naturally occurring residues will occupy the query amino acid position. In other

embodiments, such as illustrated in FIG. 2 and FIG. 6, peptide pool variations at

15   the query position were selected from a reduced set of amino acids.

Because scoring of potential sites in proteins requires a PSSM that

includes information on all naturally occurring residues, use of reduced sets

requires extrapolation of information from tested residues to residues that have

not been tested. The methods of the invention can readily be expanded to

20   include additional residues that provide data to test whether the extrapolated

results (e.g. those at the bottom of the chart in FIG. 5) are valid.

For example, FIGs. 16 and 17 show scores for the P+1 position of PKC

theta using test set 1 (see also FIG. 2) and a test set 2 that is identical in sequence

except that it includes 4 additional query residues and was synthesized several

25   months after test set 1. The two sets were tested in two different experiments

that were performed several months apart. Nonetheless, the table and graph in

FIGs. 16 and 17 show that the scores for the residues tested are in very good

agreement. The results also showed generally adequate agreement between

values extrapolated for untested residues and the values subsequently

30   experimentally determined for those residues. For example, the Log Score for

methionine at position P+1 was extrapolated to be 0.7 and experimentally shown

to be 0.8. However, the experimentally determined Log Score value for tyrosine

(0.5) did differ somewhat from the extrapolated value (1.4). Because the

differences in extrapolated and experimentally determined values for tyrosine

and phenylalanine were larger than optimal, in preferred embodiments test sets include both F and Y as query residues.

EXAMPLE 3: Scoring phosphorylation sites Sequences from a PSSM and predicting best phosphorylation sites

The prior art provides a scoring system by which kinase substrate preferences can be used to make predictions about phosphorylation by the kinase (Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. 2001. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol 19:348-353). This example illustrates how that scoring approach is done and validates the methods described herein when applied to a known PKC substrate.

Methods Employed

As shown in FIG. 18, a raw total score can readily be calculated for any peptide sequence using the data in a PSSM, for example, the PSSMs provided in FIG. 5, FIG. 7, and FIG. 16. The total score was determined by adding together the PSSM score for each of the residues of the peptide. This type of calculation is illustrated in FIG. 18 for a peptide corresponding to a known PKC phosphorylation site in the protein MARCKS having the sequence KKKKKRF-S-FKKSFK (SEQ ID NO:80). The score derived was for the sequence surrounding the Ser-159 of the intact MARCKS protein. For example, because the P-7 position of MARCKS was occupied by K, a score of 0.4 from column P-7 of FIG. 7 was used. The scores for the other thirteen residues were similarly derived from columns of FIG. 5, FIG. 7, and FIG. 17. The fourteen scores were combined for a total score of 7.4 for the KKKKKRF-S-FKKSFK (SEQ ID NO:80) sequence in MARCKS.

The raw total scores are informative in ranking individual peptides. However, it was even more useful to estimate the relative likelihood of phosphorylation of a peptide compared to many other peptides in the human proteome (i.e. proteins encoded by human genes). Such an estimate can be conveniently represented by a percentile score. To convert a raw score for a peptide to a percentile score, a relevant set of peptide scores must first be collected and sorted. Then, the relative position of the raw total score within that ordered set is determined.

Peptide sequences were examined that surrounded 1,071,932 Ser and Thr

residues found in proteins encoded by 15651 human genes catalogued in the

human reference sequence (RefSeq) collection maintained by the National

Center for Biotechnology Information. The sequence of each protein was

5   scanned to identity each residue that could be phosphorylated on Ser or Thr..

The sequence surrounding each of these sites was used to calculate a raw score

for that site for each PSSM. The distribution of scores was determined, as

illustrated, for example, in FIG. 19 for the PKC-theta PSSM. The median score

for all these proteins was -0.9.

10   From this distribution, a percentile score was determined for any given

raw score. For example, a raw score of > 2.8 corresponds to the top 5 percentile

and a raw score of >6.2 corresponds to the top 0.2 percentile of sites likely to be

phosphorylated by a selected kinase. Using this distribution, each score can be

assigned a percentile. For example, a raw score of 7.4 for the KKKKKRF-S-

15   FKKSFK (SEQ ID NO:80) sequence in MARCKS corresponds to the 0.04

percentile. Such a low percentile indicates that the KKKKKRF-S-FKKSFK

(SEQ ID NO:80) sequence in MARCKS is amongst the best candidate substrates

for PKC. Therefore, this kind of finding indicates that using the PSSM provided

by FIG. 5, FIG. 7, and FIG. 17, one of skill in the art can predict which sequence

20   within which protein is particularly likely be phosphorylated by PKC-theta.

In another embodiment, the invention provides methods for identifying

which sites in a protein of interest are likely to be phosphorylated by a particular

kinase, such as PKC-theta. FIG. 20 illustrates such an analysis for the thirty nine

Ser and Thr residues in the protein MARCKS. The panel on the left shows the

25   percentile score for each of the thirty nine residues. There is only one region of

the MARCKS protein in which PKC phosphorylation sites are likely located.

The panel on the right shows a portion of the analysis corresponding to this most

likely region. Each row shows a candidate site, together with information on the

position of the candidate site, and percentile predictions for phosphorylation at

30   the candidate position by three kinases studied: PKC-theta, AKT1, and PKA. As

shown in FIG. 20, two very strong candidate sites exist for PKC-theta at P0

positions 159 and 163 (percentile< 0.2). The values for AKT1 and PKA suggest

there are much less likely to be sites for phosphorylation by those kinases.

These sites are precisely the two sites known to be physiologically relevant PKC

phosphorylation sites in MARCKS. This kind of validation has been reproduced in a number of other molecules with known PKC phosphorylation sites, such as alpha-, beta-, gamma-adducins, and GAP-43.

'5   **EXAMPLE 4: Identification of *in vitro* phosphorylation sites for PKC**

Many peptides that are good substrates for PKC enzymes were identified using the methods of the invention. For example, Tables 4 and 5 provide a listing of peptides identified as potentially useful kinase substrates. The locuslink identifier (NCBI) for the gene, the gene symbol and the peptide

10   sequence, together with results for results for phosphorylation by up to seven different kinases are provided Tables 4 and 5. Five PKC isoforms were tested using the methods described herein (see, e.g. Example 1): one classical PKC isoform (PKC-alpha), three "novel" PKC isoforms (PKC-epsilon, PKC-delta and PKC-theta) and one atypical PKC isoform (PKC-zeta). The data provided in

15   Tables 4 and 5 show that novel and classical PKCs exhibit similar phosphorylation site preferences. In contrast to the general similarity of the substrates selected by the four classical PKC isoforms tested (PKC-alpha, PKC-epsilon, PKC-delta and PKC-theta), a more distant PKC isoform (PKC-zeta) and two other kinases in the same superfamily (AGC) show rather different patterns

20   of phosphorylation. Note that Table 5 includes data for two different concentrations of substrate peptide during the assay (10µM and 1µM). Results are substantially similar at those two concentrations, indicating that these findings on specificity are of general relevance and pertain to phosphorylation over a broad range of substrate concentrations.

Table 4

Identification of additional PKC substrates.

PKC isoforms alpha, epsilon, delta and theta have similar specificity.

LocusLink	Name or Gene Symbol	P0 Position	SEQ ID NO:	Sequence	Average	PKC alpha	PKC-epsilon	PKC delta	PKC theta	PKC-zeta	PKA	AKT1
4296	MLK3	477	76	HVRRRRGTF KRSKLRARD	91	62	100	100	100	41	69	100
8525	DGKZ	265	77	KKKKRASFK RKSSKKG	80	100	71	76	74	60	4	8
5341	PLEK	0	78	KFARKSTRRS IRLPE	63	69	84	52	47	100	5	2
9162	DGKI	345	79	NRKKKRTSF KRKA	60	50	92	66	33	18	3	7
4082	MARCKS	159	80	KKKKKRFSF KKSFKL	56	65	71	24	63	16	3	6
5339	PLEC1	0	81	KRERKTSSKS SVRKRR	56	62	59		46	10	2	9
9828	p164-RhoGEF	369	82	PRLRRGSKK RPAR	56	61	72	40	49	38	10	9
1128	CHRM1	451	83	RKIPKRPGSV HRTPSRQ	55	41	47	38	92	4	2	6
2561	GABRB2	472			53	39	52	34	87	19	63	10

1662.009WO1

LocusLink	Name or Gene Symbol	P0 Position	SEQ ID NO:	Sequence	Average	PKC alpha	PKC-epsilon	PKC delta	PKC theta	PKC-zeta	PKA	AKT1
			84	QKKSRLRRR ASQLKI								
5578	PRKCA	25	85	RFARKGSLR QKNV	52	42	86	34	46	56	81	10
3757	KCNH2	890	86	RQRKRKLSFR RRTDKD	48	47	63	42	38	28	87	39
94121	SYTL4	414	87	RQGKRKTSIKRD TVNPL	47	46	31		65	3	64	7
65108	MACMA RCKS	104	88	KKPFKLSGLSFK RNRKE	43	48	44		38	4	3	10
55357	PARIS1	449	89	EYLERRASRRRA V	41	37	45	20	60	14	4	1
286	ANK1	68	90	AQIVKRASLKRG KQ	40	40	49	32	38	5	3	3
5587	PRKCM	437	91	VHYTSKDTLRK RHYWR	40	60	49		10	12	3	5
395	ARHGAP 6	257	92	DGQKRKKSLRK KLD	38	36	51	17	47	2	18	2
9266	PSCD2	392	93	AARKKRISVKK KQEQ	37	34	40	35	40	1	1	7
2081	ERN1	724	94	KLAVGRHSFSRR SGV	36	34	13	12	86	10	5	9
119	ADD2	713		KKKFRTPSFLKK	33	34	31	25	42	5	2	9

1662.009WO1

LocusLink	Name or Gene Symbol	P0 Position	SEQ ID NO:	Sequence	Average	PKC alpha	PKC-epsilon	PKC delta	PKC theta	PKC-zeta	PKA	AKT1
			95	SKK								
775	CACNA1C	1898	96	RGFLRSASLGRRASFHLE	33	40	41	18	31	22	86	32
434	ASIP	78	97	KKRSSKKEASMKKVVRP	28	37	16		31	1	1	10
393	ARHGAP4	217	98	AGPLRKSSLKKGGRL	27	25	34	17	22	4	21	7
9590	AKAP12	311	99	AGWRKKTSFRKPKED	27	37	27	17	25	10	1	6
9020	MAP3K14	140	100	WKGKRRSKARKKRK	26	14	13	22	56	8	2	3
4687	NCF1	0	101	GAPPRRSSIRNAH	23	32	19	13	29	3	100	12
4763	NF1	2813	102	AGSFKRNSIKKIV	23	35	14	14	28	2	5	6
4607	MYBPC3	0	103	LLKKRDSFRTPRDSKLE	22	32	21	12	24	8	3	7
8436	SDPR	235	104	EKIKRSSLKKVDSLKK	22	34	17	10	25	3	1	2
9148	NEURL	238	105	ALRRPSLRREADD	21	26	29	9	21	4	84	12
1385	CREB1	0	106	EILSRRPSYRKILND	21	23	31	9	19	15	30	12
94274	PPP1R1	38			19	18	7	10	40	1	2	3

71

1662.009WO1

LocusLink	Name or Gene Symbol	P0 Position	SEQ ID NO:	Sequence	Average	PKC alpha	PKC-epsilon	PKC delta	PKC theta	PKC-zeta	PKA	AKT1
	4A		107	QKRHARVTVKYDRRE								
3985	LIMK2	473	108	KATTKKRTLRKNDRK	16	19	8	6	32	1	0	1
8013	NR4A3	366	109	KEVVRTDSLKGRRGR	16	21	26	7	9	2	1	2
57731	SPTBN4	2555	110	EGGDRRASGRRK	15	30	4	5	22	3	1	1
	MARCKS		111	KKKRFSFKKSFKLSGFSFKK	14	11	12	16	18	9	3	3
10969	EBNA1BP2	289	112	KRPGKKGSNKRPGKR	12	10	11	8	17	1	0	3
54986	FLJ205744	174	113	GENVLKKSMKSRVKG	10	16	10	4	10	1	0	9

72

1662.009WO1

Table 5
Identification of additional substrates for PKC.
Specificities of PKC-theta, -delta, -epsilon and -alpha are similar.

Sequence	SEQ ID NO:	Locus Link ID	Std Name	P0Range	average	PKC theta		PKC-delta		epsilon	alpha		zeta	
					Peptide conc. (uM)	10	1	10	1	1	10	1	10	1
KKKKRASFKRKSSKKG	114	8525;	Dag kinase zeta	265-265;	95	86	100	100	80	100	100	100	51	45
NRKKKRTSFKRKA	115	9162;	Dag kinase iota	344-344;	80	77	89	99	94	91	56	55	10	11
EEGTFRSSIRRLSTRRR	116	5348;	Phosphoholem man	79-79;	71	78	83	41	75	92	57	70	10	100
NRKKKRTSFKRKA	117	9162;	dag kinase iota	344-344;	70	78	84	79	100	67	39	40	11	20
RPQNTLKASKKKRASFKRK	118	8525;	dag kinase zeta	254-254;	64	52	67	60	73	89	60	46	57	46
KKRFSFKKSFKLSGFSFKKN	119	4082;	MARCKS	159-159;	61	28	69	16	42	81	98	94	20	20
AKRRRLSSLRASTSK	120	6194;	ribosomal protein S6	235-235;	52	51	51	32	64	56	58	55	34	43
LRRRSLRRSNSISKSPGP	121	3985;	LIMK-2	283-283;262-262;	50	57	51	53	80	49	29	33	56	60
RAITSTLASSFKRRR	122	2902;	NMDAR1	884-884;	49	42	66	24	55	72	44	43	17	20
KKRFSFKKSFKLSGFSFKKN	123	4082;	MARCKS	159-159;	49	32	47	22	45	67	73	58	20	20

1662.009WO1

Sequence	SEQ ID NO:	Locus Link ID	Std Name	P0Range	average Peptide conc. (uM)	PKC theta		PKC-delta		epsilon	alpha		zeta	
					10	10	1	10	1	1	10	1	10	1
PRLIRRGSK KRPAR	124	9828;	p164-RhoGEF	922-922;	49	57	53	43	69	49	24	48	32	45
PLKEKKRE RKTSSKSS VRKR	125	5339;	Plectin	4157-4157;	48	100	39	36	46	48	31	34	12	13
KAIKAIEG GQKFARKS TRRS	126	5341;	pleckstrin 1	113-113;	46	45	46	43	47	69	34	36	56	46
SQVQKQRS AGSFKRNSI KKI	127	4763;	NF1	2798-2798;	43	58	45	33	52	32	39	40	19	19
QQVDRERP HVRRRGT FKRS	128	4296;	MLK3	477-477;	42	59	41	57	46	32	33	25	9	14
VQRHRSM RKTFARYL SFRRD	129	4171;	MCM2;	801-801;	40	25	48	14	32	65	37	58	31	51
EYLERRAS RRRAV	130	55357;	PARIS 1	443-443;	39	44	38	20	46	46	38	38	39	36
WKGKRRS KARKKRK	131	9020;	NIK	140-140;	38	35	53	29	38	43	22	49	6	7
GFLNEPLSS KSQRRKSL KLK	132	57082;	AF15q 14;	1059-1059;1059;1059-1059;1085-1085;408-408;	38	40	34	47	41	47	30	24	61	19
LEKRGMLG KRPRRKSS RRKK	133	3797;	kinesin 3C		37	53	46	52	46	31	16	17	30	27
RSRSRSRS KSKDKRKS	134	6429;	splicing factor,	342-342;	34	16	36	16	44	32	39	54	5	5

74

1662.009WO1

Sequence	SEQ ID NO:	Locus Link ID	Std Name	P0Range	average Peptide conc. (uM)	PKC theta [10]	PKC theta [1]	PKC-delta [10]	PKC-delta [1]	epsilon [1]	alpha [10]	alpha [1]	zeta [10]	zeta [1]
RKRS			arginine/serine-rich 4		1	10	1	10	1	1	10	1	10	1
KKKFRTPSFLKKSKK	135	119;	beta adducin	711-711;	33	36	30	25	40	27	19	56	11	13
RARRDSLKKIEIW	136	9101;	ubiquitin specific protease 8	994-994;	32	28	42	23	41	33	28	26	6	9
PSKSPSKKKKKFRTPSFLKK	137	119;	beta adducin	699-699;	31	19	40	16	38	32	38	34	4	5
EYLERRASRRRAV	138	55357;	PARIS 1	443-443;	31	39	34	28	49	27	20	17	31	61
RPTPGDGEKRSRIKKSKKRK	139	79142;	MGC2941;	205-205;	30	17	31	24	31	24	44	39	1	1
TELEGGFSRQRKRKLSFRRR	140	3757;	HERG	875-875;	30	35	25	49	44	27	15	12	24	49
VTDSQKRREILSRRPSYRKI	141	1385;	CREB	105-105;	29	28	29	21	43	31	25	24	26	78
ERHVAQKKSRLRRRASQLKI	142	2561;	GABA A receptor beta 2	465-465;	28	37	28	36	30	28	21	16	22	37
VRYTPYTISPYNRKGSFRKQ	143	56000;	nuclear RNA export	66-66;	28	27	25	32	33	21	19	38	22	79

1662.009WO1

Sequence	SEQ ID NO:	Locus Link ID	Std Name	PORange	average Peptide conc. (uM)	PKC theta	PKC theta	PKC-delta	PKC-delta	epsilon	alpha	alpha	zeta	zeta
						10	1	10	1	1	10	1	10	1
LSSMFGTLPRKSRKGSVRKQ	144	9595;	factor 3 PSCDBP	322-322;	28	22	35	28	35	24	20	31	54	16
ISDFGLAKKLAVGRHSFSRR	145	2081;	ERN1	710-710;	27	25	27	23	34	26	30	21	61	25
QAQRQIKRGAPPRRSSIRNA	146	4687;	p47phox	303-303;	27	34	33	14	33	28	24	20	56	32
RDIRQSPKRGFLRSASLGRR	147	775;	calcium channel, voltage-dependent, L type, alpha	1924-1924;		69	20	13	27	26	16	13	34	28
RELEQLKAEYLERRASRRRA	148	55357;	PARIS 1	443-443;	26	34	34	14	27	33	22	19	20	20
RVVQSVKHTKRKSSTVMK	149	5587;	PKD1	412-412;	25	14	13	17	20	27	52	33	5	3
VDPFYEMLAARKKRISVKKK	150	9266;	cytohesin-2	381-381;	24	35	35	20	34	20	11	12	6	3
PQNSLKASNRKKKRTS	151	9162;	dag kinase iota	333-333;	24	19	25	37	27	24	16	17	28	15

1662.009WO1

Sequence	SEQ ID NO:	Locus Link ID	Std Name	P0Range	average Peptide conc. (uM)	PKC theta (10)	PKC theta (1)	PKC-delta (10)	PKC-delta (1)	epsilon (10)	epsilon (1)	alpha (10)	alpha (1)	zeta (10)	zeta (1)
FKRK						10	1	10	1	10	1	10	1	10	1
DLIEGRKG AQIVKRAS LKRG	152	286;	ankyrin R	68-68;	23	30	31	18	35	25		12	13	9	4
TYLLPDKS RQGKRKTS IKRD	153	94121;	slp4	399-399;	23	18	22	26	32	24		22	17	13	8
KKFFTQGW AGWRKKT SFRKP	154	9590;	gravin	301-301; 203-203;	22	21	34	14	26	29		12	16	23	27
RWDKRRW RKIPKRPGS VHRT	155	1128;	M1 muscarinic receptor	451-451;	21	26	23	23	30	26		12	10	6	19
SAQITIPKD GQKRKKSL RKK	156	395;	ARHGAP6	242-242;	20	23	18	31	29	18		11	9	44	15
PSPSNETPK KKKKRFSF KKS	157	4082;	MARCKS	145-145;	20	23	20	35	27	12		11	10	6	3
VQMTWSY PDEKNKRA SVRRR	158	2321;	flt1	265-265;	19	23	19	33	18	11		11	16	72	11
LYARLARA YRRSQRAS FKRA	159	2837;	Urotensin-2 receptor	231-231;	19	17	21	12	15	25		19	21	41	25
PFEVVWYK DKRQLRSS KKYK	160	7273;	titin	6478-6478;	18	25	19	31	25	13		7	9	57	15

1662.009WO1

Sequence	SEQ ID NO:	Locus Link ID	Std Name	P0Range	average Peptide conc. (uM)	PKC theta		PKC-delta		epsilon		alpha		zeta	
KYKAFIRIP IPTRRHTFR RQ	161	5337;	PLD1	133-133;		10	1	10	1	1		10	1	10	1
KKKFSFKK PFKLSGLSF KRN	162	65108;	MacM ARCK S	93-93;	18	23	19	14	22	21		8	22	9	15
PPRTPGWH QLQPRRVS FRGE	163	9088;	Myt1 kinase	71-71;	17	13	14	13	21	19		19	17	4	7
TEGKMAR VAWKGKR RSKARK	164	9020;	NIK	125-125;	16	11	14	21	13	14		10	30	26	7
TEEKSKKR KKHRKNS RKHK	165	9360;	cyclo philin G	223-223;	16	20	24	26	18	12		6	6	2	3
MAQIERGE ARIQRRISI KKA	166	6594;8 467;	SMAR CA5	931-931;910-910;916-916;	15	14	16	19	25	12		10	15	2	1
GLPAPGED KSIYRRGS RRWR	167	5590;	pkc-zeta	113-113;113-113;	15	13	12	29	13	10	6	7	27	37	5

78

Quantitative analysis of correlations between phosphorylation of the same substrate by different kinases is shown in FIG. 21. Such analysis confirms the conclusions that the novel and classical PKC isoforms are very similar in specificity, that there is greater divergence of the atypical PKC isoform PKC-

5   zeta, and that the other kinases of the same superfamily (AGC) are even more divergent in specificity.

Results in Table 2, Table 3, Table 4 and Table 5 demonstrate phosphorylation by PKC of many of the peptides. As validated herein, the methods of the invention predict that Ser and Thr residues within those peptides

10   are the preferred sites of phosphorylation. Table 6 lists sequences of peptides in which pSer and pThr are present at positions corresponding to preferred PKC phosphorylation sites in peptides phosphorylated by PKC. Phosphopeptides included in Table 6 are only those corresponding to peptides whose efficiency of phosphorylation by PKC is greater than or equal to 10% of the best substrate.

15   Such a cutoff is relatively stringent. It is more rigorous than many previous methods in which the magnitude of phosphorylation is not compared with reference positives.

TABLE 6. Sequence of phosphopeptides
corresponding to preferred sites of PKC phosphorylation

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
301	202	absent in melanoma 1	RSGRRRG-pS-QKSTDS	0.0	N
302	286	ankyrin R	AQIVKRA-pS-LKRGKQ	0.3	
303	695	BTK	FERGRRG-pS-KKGSID	0.2	
304	1105	CHD1	SEGRRSR-pS-RRYSGS	0.1	
305	1455	casein kinase I gamma 2	FKRRKRK-pS-LQRHK-	0.1	
306	1612	DAP-kinase 1	IKKRRTK-pS-SRRGVS	0.0	
307	1612	DAP-kinase 1	KKRRTKS-pS-RRGVSR	0.2	
308	1794	DOCK2	PEVKLRR-pS-KKRTKR	0.1	
309	1901	S1P1 receptor	YSLVRTR-pS-RRLTFR	0.1	
310	2870	GRK6	GGNRKGK-pS-KKWRQM	0.5	N
311	3985	LIMK-2	---LRRR-pS-LRRSNS	0.0	
312	4033	JAW1	RFSRRSS-pS-WRILGS	0.6	
313	4296	MLK3	RRGTFKR-pS-KLRARD	0.8	
314	4296	MLK3	HVRRRRG-pT-FKRSKL	0.0	
315	4542	myosin IF	KKERRRN-pS-INRNFV	0.0	
316	4820	NKTR	TSSYRSR-pS-YSRSRS	0.7	
317	5128	PCTK2	KKFKRRL-pS-LTLRGS	0.1	N
318	5339	Plectin	RKTSSKS-pS-VRKRR-	0.5	
319	5734	prostaglandin E receptor 4	SDFRRRR-pS-FRRIAG	0.0	
320	5777	SHP-1	DKEKSKG-pS-LKRK--	2.0	C
321	5778	HePTP	RALSFRQ-pT-SWLS--	2.0	
322	5778	HePTP	EQQRRAL-pS-FRQTSW	3.0	N

80

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
323	7074	**TIAM1**	QAMSRSA-pS-KRRSRF	0.6	
324	9221	**nucleolar phosphoprotein p130**	KTKKKRG-pS-YRGGSI	0.5	C
326	9360	**cyclophilin G**	KKKHRKN-pS-RKHK--	0.0	
327	9595	**PSCDBP**	FGTLPRK-pS-RKGSVR	0.2	
328	9595	**PSCDBP**	PRKSRKG-pS-VRKQ--	0.0	
329	9595	**PSCDBP**	SSSRRNR-pS-ISN---	0.3	
330	9595	**PSCDBP**	DFLRRSS-pS-RRNRSI	0.3	
331	23031	**MAST3**	--RMARR-pS-KRSRRR	0.2	
332	23031	**MAST3**	ETQDRRK-pS-LFKKIS	0.4	
333	23031	**MAST3**	MARRSKR-pS-RRRETQ	0.2	
334	25836	**IDN3**	RRRSQRI-pS-QRIT-	0.0	
335	25836	**IDN3**	SGVRRRR-pS-QRISQR	0.0	
336	26191	**Lyp**	VILRPSK-pS-VKLRSP	0.6	
337	65125	**WNK1**	RRRRPTK-pS-KGSKSS	1.0	
338	65125	**WNK1**	SGRRRRP-pT-KSKGSK	0.0	
339	65125	**WNK1**	RKSVRSR-pS-RHE--	0.6	
340	65125	**WNK1**	TKRHYRK-pS-VRSRSR	0.0	
341	393	**ARHGAP4**	AGPLRKS-pS-LKKGGR	0.3	
342	409	**beta-arrestin2**	EKSHKRN-pS-VRLVIR	0.5	
343	119	**adducin gamma**	TPSFLKK-pS-KK----	2.0	
344	202	**absent in melanoma 1**	-----RR-pS-GRRRGS	0.5	N
345	395	**ARHGAP6**	DGQKRKK-pS-LRKKLD	0.1	
346	672	**BRCA1**	NRLRRKS-pS-TRHIHA	0.1	
347	672	**BRCA1**	-NRLRRK-pS-STRHIH	0.1	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
349	1105	CHD1	-GSEGRR-pS-RSRRYS	0.7	
350	1196	CLK2	RRRRRSR-pT-FSRSSS	0.0	C
351	1196	CLK2	RRRSRTF-pS-RSSS-	1.0	C
352	1198	CLK3	YRWKRRR-pS-YSREHE	0.1	N
353	1794	DOCK2	LRRSKKR-pT-KRSS--	0.9	
354	2081	ERN1	KLAVGRH-pS-FSRRSG	1.0	
355	2081	ERN1	AVGRHSF-pS-RR---	4.0	
356	2305	forkhead (Drosophila)-like 16	-RERRER-pS-RSRRKQ	0.0	
357	2305	forkhead (Drosophila)-like 16	ERRERSR-pS-RRKQHL	0.4	
358	3797	kinesin 3C	KRPRRKS-pS-RRKK--	0.0	
359	3797	kinesin 3C	GKRPRRK-pS-SRRKK-	0.0	
360	3985	LIMK-2	KATTKKR-pT-LRKNDR	1.0	
361	3985	LIMK-2	RRRSLRR-pS-NSISKS	0.5	
362	3985.	LIMK-2	RSLRRSN-pS-ISKSPG	0.1	
363	4033	JAW1	DRFSRRS-pS-SWRILG	3.0	
364	4033	JAW1	-DRFSRR-pS-SSWRIL	3.0	
365	4171	MCM2;	--VQRHR-pS-MRKTFA	0.0	
366	4763	NF1	AGSFKRN-pS-IKKIV-	0.5	C
367	4820	NKTR	SYRSRSY-pS-RSRSRG	2.0	
368	4820	NKTR	RSRSYSR-pS-RSRG--	1.0	
369	4863	NPAT	RASSRST-pT-KKR---	1.0	
370	4863	NPAT	FRASSRS-pT-TKKR--	1.0	
371	5128	PCTK2	FKRRLSL-pT-LRGSQT	1.0	N

82

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
372	5339	Plectin	--KRERK-pT-SSKSSV	1.0	
373	5339	Plectin	KKRERKT-pS-SKSSVR	1.0	
374	5587	PKD1	KHTKRKS-pS-TVMK--	0.3	
375	5587	PKD1	VHYTSKD-pT-LRKRHY	3.0	
376	5590	pkc-zeta	KSIYRRG-pS-RRWR--	0.0	
377	6840	Supervillin	NVMKRKF-pS-LRAAEF	0.5	
378	7074	TIAM1	RSASKRR-pS-RFSS--	2.0	
379	8436	serum deprivation response;	-EKIKRS-pS-LKKVDS	3.0	
380	8915	BCL10	EISCRTS-pS-RKRAGK	4.0	
381	9020	NIK	-WKGKRR-pS-KARKKR	0.8	
382	9101	ubiquitin specific protease 8	--RARRD-pS-LKKIEI	1.0	
383	9148	neurlized-like	--ALRRP-pS-LRREAD	0.5	
384	9162	dag kinase iota	-NRKKKR-pT-SFKRKA	0.6	
385	9595	PSCDBP	DDFLRRS-pS-SRRNRS	1.0	
386	9828	p164-RhoGEF	PRLIRRG-pS-KKRPAR	0.0	
387	10123	ADP-ribosylation factor-like 7 = ARL7	MILKRRK-pS-LKQK--	0.0	
388	10969	EBNA1BP2	KRPGKKG-pS-NKRPGK	1.0	C
389	23227	MAST4	MVRRSKK-pS-KKKESL	0.5	
390	23227	MAST4	--RMVRR-pS-KKSKKK	0.2	
391	25836	IDN3	EVSRPRK-pS-RKRVDS	0.4	
392	25865	PKD2	ARIIGEK-pS-FRRSVV	0.2	
393	26191	Lyp	-SVILRP-pS-KSVKLR	0.9	
394	55357	PARIS1	EYLERRA-pS-RRRAV-	0.2	C

83

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
395	55672	FLJ20719	KKRRGRR-pS-TKKRRR	0.0	
396	55672	FLJ20719	KRRGRRS-pT-KKRRRR	0.0	
397	57082	AF15q14;	SKSQRRK-pS-LKLK--	0.0	
398	57731	spectrin, beta, non-erythrocytic 4	EGGDRRA-pS-GRRK--	0.9	C
399	65125	WNK1	EYRRRRH-pT-MDKDSR	0.4	
400	672	BRCA1	RLRRKSS-pT-RHIHAL	1.0	
402	1196	CLK2	-RRRRRR-pS-RTFSRS	0.0	C
403	1196	CLK2	RSRTFSR-pS-SSMK--	2.0	
404	1196	CLK2	RTFSRSS-pS-MK----	2.0	
405	1198	CLK3	--------pS-YRWKRR	2.0	
406	1198	CLK3	WKRRRSY-pS-REHEGR	2.0	N
407	1612	DAP-kinase 1	-FIKKRR-pT-KSSRRG	1.0	
408	1612	DAP-kinase 1	KSSRRGV-pS-RE---	1.0	
409	1794	DOCK2	SKKRTKR-pS-S-----	2.0	
410	2081	ERN1	RHSFSRR-pS-GV---	4.0	
412	2837	Urotensin-2 receptor	YRRSQRA-pS-FKRA--	0.0	
413	2837	Urotensin-2 receptor	LARAYRR-pS-QRASFK	0.1	
414	3985	LIMK-2	-----KA-pT-TKKRTL	2.0	
415	3985	LIMK-2	----KAT-pT-KKRTLR	4.0	
416	4171	MCM2;	RHRSMRK-pT-FARYLS	2.0	
417	4171	MCM2;	KTFARYL-pS-FRRD--	2.0	
418	4763	NF1	QKQRSAG-pS-FKRNSI	1.0	C
419	4820	NKTR	RSYSRSR-pS-RG----	5.0	
420	4863	NPAT	NTQQFRA-pS-SRSTTK	2.0	
421	4863	NPAT	TQQFRAS-pS-RSTTKK	3.0	
422	5587	PKD1	VKHTKRK-pS-STVMK-	5.0	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
	5587	PKD1	HTKRKSS-pT-VMK---	4.0	
423					
	5587	PKD1	---RVVQ-pS-VKHTKR	1.0	
424					
	6429	SFRS4	KSKDKRK-pS-RKRS--	0.2	
425					
426	6429	SFRS4	KRKSRKR-pS-------	0.6	
	6429	SFRS4	RSRSRSK-pS-KDKRKS	0.4	
427					
	6429	SFRS4	RSRSRSR-pS-KSKDKR	0.3	
428					
429	6429	SFRS4	----RSR-pS-RSRSKS	0.6	
	6429	SFRS4	--RSRSR-pS-RSKSKD	0.6	
430					
	6594		ARIQRRI-pS-IKKA---	0.1	
431		SNF2L			
			APLRRRE-pS-MHVEQR		
432	6650	SOLH		0.0	
	7273	Titin	DKKQIRS-pS-KKYR--	2.0	
433					
	7273	Titin	KDKRQLR-pS-SKKYK-	0.7	
434					
	8436	serum deprivation response;	SSLKKVD-pS-LKK--	5.0	
435					
	8567	MADD	SVRQRRM-pS-LRDD--	1.0	
436					
			SRSRHRL-pS-RSR--		
437	8621	CDC2L5	-	0.1	
			-SSRHSR-pS-RSRHRL		
438	8621	CDC2L5		0.9	
			YSRRRSP-pS-YSRHSS		
439	8621	CDC2L5		0.3	
			SRHSRSR-pS-RHRLSR		
440	8621	CDC2L5		0.4	
			-RDRGRR-pS-RSRLRR		
441	8899	PRP4		0.1	
442	8899	PRP4	RSRLRRR-pS-RS----	0.1	
			RGGRRRR-pS-RSKVKE		
443	8899	PRP4		0.0	
			TTKKRSK-pS-RSKERT		
444	8899	PRP4		0.4	
			DRGRRSR-pS-RLRRRS		
445	8899	PRP4		0.1	
446	8899	PRP4	RLRRRSR-pS-------	0.6	
			GRRRRSR-pS-KVKEDK		
447	8899	PRP4		0.0	
448	9020	NIK	-KKRKKK-pS-	2.0	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
449	9020	NIK	SKSLAH KKRKKKS-pS-KSLAHA	1.0	
450	9088	Myt1 kinase	QLQPRRV-pS-FRGE--	1.0	
451	9221	nucleolar phosphoprotein p130	-----EK-pT-KKKRGS	1.0	
452	9221	nucleolar phosphoprotein p130	RGSYRGG-pS-ISV--	0.6	
453	9360	cyclophilin G	---TEEK-pS-KKRKKK	1.0	
454	9590	Gravin	AGWRKKT-pS-FRKP--	0.4	
455	9590	Gravin	-AGWRKK-pT-SFRKPK	0.3	
456	9595	PSCDBP	-DDFLRR-pS-SSRRNR	3.0	
457	9934	GPR105	STSVKKK-pS-SRN--	2.0	
458	9934	GPR105	TSVKKKS-pS-RN---	2.0	
459	9934	GPR105	KSSRNST-pS-VKKSS	0.3	
460	9934	GPR105	LKSSRNS-pT-SVKKS	2.0	
461	9934	GPR105	-LKSSRN-pS-TSVKKK	1.0	
462	23031	MAST3	KRSRRRE-pT-QDR--	0.1	
463	26191	Lyp	VKLRSPK-pS-------	4.0	
464	55357	PARIS1	EYLERRA-pS-RRRAV-	0.2	C
465	55762	FLJ10891;	ARPKTRI-pS-NKYR--	0.8	
466	56000	nuclear RNA export factor 3	SPYNRKG-pS-FRKQ--	0.1	
467	57468	solute carrier family 12 member 5	ITDESRG-pS-IRRK-	2.0	
468	79142	MGC2941;	PGDGEKR-pS-RIKKSK	2.0	
469	79142	MGC2941;	KRSRIKK-pS-KKRK--	0.0	
470	79877	FLJ22955;	ARLMRRN-pS-LNRK--	0.0	
471	94121	slp4	-RQGKRK-pT-	1.0	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
			SIKRDT		
472	94121	slp4	RQGKRKT-pS-IKRDTV	0.4	
473	9162	dag kinase iota	NRKKKRT-pS-FKRKA-	0.0	
571	547	ATSV/KIF1A	GTIRSKL-pS-RRRSAQ	0.1	c
572	547	ATSV/KIF1A	SKLSRRR-pS-SAQMRV-	0.4	c
573	10921	RNPS1	PGRRRHR-pS-RSSSNS	0.0	c
574	10921	RNPS1	RRRHRSR-pS-SSSNSSR	0.2	c
575	25836	IDN3	SGVRRRR-pS-QRISQR	0.0	c
576	25836	IDN3	RRRSQRI-pS-SQRIT--	0.1	c
577	1608	dag kinase gamma	FFSLRRK-pS-RSKD--	0.3	c
578	1608	dag kinase gamma	PQKSSFF-pS-SLRRKSR	2.0	c
579	27330	p90-RSK6	SSLAQRR-pS-MKKRTS	0.1	c
580	27330	p90-RSK6	RSMKKRT-pS-STGL---	1.0	c
581	9014	TAF1B	YSVKRKK-pS-RSKKVR	0.0	c
582	9014	TAF1B	VKRKKSR-pS-SKKVRRH	0.2	c
583	6712	spectrin, beta, non-erythrocytic 2	REREKRF-pS-FFKKNK	0.2	c
584	941	CD80	NERLRRE-pS-VRPV--	2.0	c
585	1455	casein kinase I gamma 2	FKRRKRK-pS-LQRHK-	0.0	c
586	6621	SNAPC4	RTRHARH-pT-RKRRRL	0.1	c
587	9939	RBM8A	RRGGRRR-pS-RSPDRR	0.3	c
588	9939	RBM8A	GGRRRSR-pS-SPDRRRR	2.0	c
589	6158	RPL28	KRKRTRP-pT-KSS--	0.5	c
590	6158	RPL28	KRTRPTK-pS-SS----	2.0	c
591	9585	MPP1	KRRLRTK-pT-AK--	0.4	c

87

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
592	9585	MPP1	IIKRRLR-pT-TKTAK--	0.5	C
593	5336	PLCG2	REKRVSN-pS-KFYS--	2.0	C
594	55762	FLJ10891	ERHHRLH-pT-GKKS--	3.0	C
595	2889	RAS-GRF2	-IKPRNI-pT-RRKTDR	1.0	C
596	2889	RAS-GRF2	RNITRRK-pT-TDREEKT	1.0	C
597	117532	TMC2	DRLGRRS-pS-SKRALK	0.1	N
598	117532	TMC2	RLGRRSS-pS-SKRALKA	0.5	N
599	11215	AKAP220	NHMKTKA-pS-VRKSFS	0.1	N
600	22899	ARHGEF15	IIRPRPPSR-pS-RAAQ	5.0	N
601	10788	IQGAP2	KRKNTRR-pS-IKLDG-	0.1	
602	10788	IQGAP2	DNLKRKN-pT-TRRSIKL	1.0	
603	1620	DBCCR1	PRWRKRM-pS-LTLKSN	0.1	
604	1620	DBCCR1	WRKRMSL-pT-TLKSNKN	3.0	
605	9595	PSCDBP	SSSRRNR-pS-IS----	0.4	
606	9595	PSCDBP	DFLRRSS-pS-SRRNRSI	0.5	
607	9656	NFBD1	----TSRA-pT-RRKTNR	0.1	
608	9656	NFBD1	SRATRRK-pT-TNRSSVK	0.1	
609	785	CACNB4	HNERARK-pS-RNRLSS	0.5	
610	785	CACNB4	RKSRNRL-pS-SSSS---	0.8	
611	9726	KIAA0296 gene product	RAYRHRG-pS-LVNHRH	0.0	
612	54221	SNTG2	GRNRRTV-pT-LRRQPV	0.1	
613	54221	SNTG2	HQGRNRR-pT-TVTLRRQ	0.2	
614	22947	DUX4	SRRPPRR-pS-RSRRPG	0.1	
615	22947	DUX4	RPPRRSR-pS-SRRPGLH	0.1	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
616	23524	SRm300	--RKARL-pS-RRSRSA	0.2	
617	23524	SRm300	KARLSRR-pS-SRSASSS	0.4	
618	8471	IRS4	HLPRGRR-pS-RRAVSV	0.4	
619	8471	IRS4	RRSRRAV-pS-SVPA---	0.7	
620	4926	NUMA	RSARRRT-pT-QI----	0.2	
621	4926	NUMA	TRSARRR-pT-TTQI---	0.4	
622	2318	gamma-filamin	TRTFTRS-pS-HTY---	2.0	
623	9656	NFBD1	RGRKNRS-pS-VKTPET	0.4	
624	9656	NFBD1	TRGRKNR-pS-SSVKTPE	0.7	
625	11214	AKAP13	TKVSRTF-pS-YIKNKM	3.0	
626	9744	centaurin beta 1	SIRPRPG-pS-LRSKPE	1.0	
627	10129	hypothetical protein CG003	ERSRHQR-pS-FSVPKK	0.1	
628	9656	NFBD1	PNRIPSR-pS-LRRTKL	0.5	
629	9656	NFBD1	PSRSLRR-pT-TKLNQ--	3.0	
630	9656	NFBD1	PKIRTRK-pS-SRMTPF	0.0	
631	4690	NCK1	KNSARKA-pS-IVKNLK	0.3	
632	4690	NCK1	----RKN-pS-SARKASI	1.0	
633	862	CBFA2T1	EKTRRSL-pT-VLRRAQ	0.8	
634	862	CBFA2T1	MVEKTRR-pS-SLTVLRR	1.0	
635	1793	DOCK1	GYTLRKK-pS-KKG---	0.5	
636	1793	DOCK1	EGWYRGY-pT-TLRKKSK	1.0	
637	8826	IQGAP1	REMKGKK-pS-KKISLK	0.1	
638	8826	IQGAP1	GKKSKKI-pS-SLKYT--	0.5	
639	926	CD8beta	AQPTKKS-pT-LKKRVA	1.0	

SEQ ID NO	Locus-Link ID		Sequence indicating site of phosphorylation	Percentile prediction for PKC-theta	N or C-term
640	926	CD8beta	TAQPTKK-pS-STLKKRV	2.0	
641	10198	MPHOSPH9	MLSLRHN-pS-RIHVRP	2.0	
642	10198	MPHOSPH9	SRIHVRP-pS-SR-----	2.0	
643	8842	CD133	VRTRIKR-pS-RKLADS	0.1	
644	8842	CD133	----QVR-pT-TRIKRSR	3.0	

EXAMPLE 5: Analysis of different kinases using the same superset

In many embodiments of the invention, the same superset of test peptides

5   can be used to study the substrate specificity of a variety of different kinase

enzymes. The anchor residue(s) and phosphorylatable residue in a test set (or

superset, or collection) of peptides must be appropriate to the particular kinase

whose specificity is being analyzed. However, a wide diversity of peptide

sequences is available in the test sets, supersets, or collections of peptides

10   provided by the invention. It is also fortunate that the results obtained to date

indicate that there is sufficient similarity between the substrate specificities of

different kinases that a single set (or superset, or collection) of peptide pools can

be used to study the specificity of different kinases. Hence, for example, kinases

of the protein kinase C family are sufficiently closely related that successful

15   studies with other members of this family can be performed on the same or

similar test sets of peptides. This was shown by studies that where one or both

of the supersets of peptides designed for PKC were successfully used to analyze

related kinases such as PKC-zeta, Protein Kinase A (PKA) and Protein Kinase G

(PKG). See FIG. 22 and FIG. 25.

20   FIG. 22 shows PSSM Logos for PKC-zeta and PKA derived by

analyzing those kinases with the same peptide supersets used for analysis of

PKC-theta. Because the sequence of PKC-zeta is similar to the PKC-theta

sequence, PKC-zeta was expected to have fundamental similarities in substrate

specificity. Those expectations were confirmed by the PSSM Logo

25   representation of the data. One of the most prominent differences between PKC-

theta and PKC-zeta was the preference for a hydrophobic amino acid (e.g.,

phenylalanine, F) at P-5. This characteristic preference of PKC-zeta was

confirmed using the methods of the invention and was further validated by

previous tests (Nishikawa K, Toker A, Johannes FJ, Songyang Z, Cantley LC.

5   1997. J Biol Chem 272:952-960). Similarly, PKA has a strong preference for

positively charged residues in positions P-2 and P-3 (FIG. 22), as previously

shown by Kreegipuu A, Blom N, Brunak S, Jarv J. 1998. Statistical analysis of

protein kinase specificity determinants. FEBS Lett 430:45-50.)

Predictions were made as to which amino acids would occupy what

10   positions in the phosphorylation substrate recognized by PKC-zeta. These

predictions were then tested by measuring PKC-zeta mediated phosphorylation

of the same set of proteomic peptides that were tested for PKC-theta. The results

for this testing are shown in FIG. 23 (panel a) and demonstrate that the PKC-zeta

prediction was excellent. The quality of the prediction was affirmed by the

15   comparison with the results of predictions by the Scansite for PKC-zeta (FIG.

23, panel b). Problems with the Scansite prediction were evident from the

finding that the best peptide has a score of >4th percentile and several other of

the better substrates also have scores >4th percentile.

Given the similarity between the PSSM Logo for PKC-zeta and PKC-

20   theta, it was possible that the good results for PKC-zeta and PKC-theta are

redundant, and that nothing new has been learned from PKC-zeta. That

possibility was addressed in two ways. First, the data were checked to ascertain

whether PKC-delta/theta and PKC-zeta were equivalent in their phosphorylation

of the set of proteomic peptides. Results in FIG. 23 (panel c) show that although

25   there was a general correlation between the phosphorylation patterns of those

different kinases, there were also substantial differences. Therefore, an analysis

was performed on whether the PKC-zeta prediction would satisfactorily predict

phosphorylation by PKC-delta. The results in FIG. 23 (panel d) demonstrate that

PKC-zeta predictions would not. Thus predictions from the PKC-zeta PSSM

30   predict well phosphorylation by PKC-zeta but not PKC-theta while predictions

from the PKC-theta PSSM predict well phosphorylation by PKC-theta (and

PKC-delta). These findings strongly validate the high degree of specificity

provided by the methods of the invention.

Further investigations were performed to ascertain what residues may account for differences between substrates in the predicted phosphorylation by PKC-theta and PKC-zeta. FIG. 24 provides a detailed analysis of the scoring for the six substrates whose behavior contributed most to the mismatch in FIG. 23,

5   panel d (and corresponding match in FIG. 23, panel a). Scoring for those peptides with the PKC-theta and PKC-zeta predictions were tabulated. Residues that showed the biggest improvement in score with PKC-zeta relative to PKC-theta were identified (difference >0.5) and are underlined. Better recognition by PKC-delta could be due to a favorable residue for PKC-delta recognition that is

10   less favorable for PKC-zeta recognition (referred to herein as "control by favorable residue"), or to neutral residue for PKC-delta recognition being unfavorable for PKC-zeta recognition ("control by unfavorable residue"). The results indicate that much of the poorer recognition by PKC-zeta was due to at least one unfavorable residue. For example, the six biggest changes in score for

15   each peptide have been boxed in black in FIG. 24. Five of those six changes are from a residue slightly unfavorable for PKC-theta to a residue very unfavorable for PKC-zeta. This is best illustrated by peptides 2 and 3, which have a proline at -5 that was slightly unfavorable for PKC-theta and very unfavorable for PKC-zeta. The strongly disfavored proline at -5 for PKC-zeta (but not for PKC-theta)

20   can be seen in FIG. 22. This principle is similarly illustrated by the peptide 1, which has an isoleucine at P-1 (predicted as being disfavored based on the results for leucine with PKC-zeta, FIG. 22) and peptide 5, which has an W at P-5 (strongly disfavored by PKC-zeta, FIG. 22).

Control of kinase specificity by unfavorable residue(s) was also strongly

25   suggested by the findings that PKA, PKC-theta and PKC-zeta all strongly disfavor proline at P+1 (FIG. 22). This contrasts sharply with the preferences of another major class of kinase, the proline-directed kinase, for which a Proline at P+1 is a critical residue. Thus, an important part of the reciprocal specificity between the basophilic kinases and the proline-directed kinases (such as CDK1)

30   is that proline at P+1 was disfavored by the former and favored by the latter. Thus, "control by unfavorable residue" appears to be a major element in kinase specificity. This is important, because the methods of the invention can be very accurate at quantifying unfavorable recognition. Many of the prior art techniques may not be ideal for determining strength of unfavorable recognition;

for example, the methods disclosed in U.S. Patent 6,004,757 may be limited in doing so by reason of limitations in amino-acid sequencing.

EXAMPLE 6: Analysis of mutant kinases

5   In another embodiment, the methods of the invention can be used to analyze the substrate specificity of mutant kinases. A major strategy for analyzing protein structure and function involves deriving mutant constructs, expressing them, and determining how the mutation influences the function and/or specificity of the resulting mutant protein. Given the previous difficulty

10   in assessing kinase specificity, there have been no prior studies that systematically analyze the specificities of mutant kinases. However, the methods of the invention can be used for this purpose.

For example, more than ten mutant constructs of PKC-theta have been made and analyzed by the inventor using the present methods to ascertain what

15   types of specificity changes occur. Results of some of the more informative constructs are shown as PSSM logos in FIG. 26. Because only changes in substrate specificity were assessed and not changes in auto-inhibition resulting from altered binding of pseudo-substrate, the parental construct PKC-theta was used that had been previously mutated to a constitutively active form by

20   mutating the pseudo-substrate (A148E), shown in FIG. 26. Results are shown for four constructs in which acidic residue in the catalytic cleft has been mutated (FIG. 26).

The most striking finding amongst the constructs studied was deviation of construct D465A from the overall pattern of substrate specificities shared by

25   wild type PKC-theta (FIG. A), constitutive active A148E (FIG. 26) and the three other mutant constructs derived from constitutive active A148E (D544A, D508A, E571I, FIG. 26). The differences observed in D465A specificity compared to other PKC-theta enzymes are: 1) the shapes of the PSSM Logo (i.e. relative height of individual columns) and 2) the general position of individual

30   residues in particular columns.

Regarding the shape of the PSSM Logo, a feature absolutely conserved amongst constructs other than D465A was that the P+2 position was always the tallest. Usually the P+1 position was the second tallest and there was wobble as to which of the other positions was third tallest. However, mutant D465A was

strikingly different. Position P+2 of the preferred substrate for the D465A
mutant has dropped from the most prominent to one of the three least prominent
and the P+1 position has likewise dropped in prominence. Taken together these
data indicate that the D465A mutant has a marked reduction in reliance on the

5   usual C-terminal residues that typically guide substrate specificity in all other
kinase constructs.

A detailed understanding of kinase specificity requires understanding of
the residues favored at each position. PSSM Logos (FIG. 26) also reveal that the
strong preferences and lack of preferences of the wild type construct for residues

10   at particular positions was typically conserved amongst most mutant kinase
constructs. These generally include: 1) a preference for basic residues at each
position; 2) an absolute preference for a hydrophobic residue that exceeds the
preference for basic residues at the P+1 position (and occasionally P+3); 3) a
strong disfavor for aspartic acid ('D') at most positions; 4) a strong dislike for

15   hydrophobic residues at P-2; and 4) a strong disfavor for proline ('P') in a C-
terminal position. As with the overall shape, D465A was also an outlier with
regard to these preferences and disfavors. Note particularly the moderation, or
reversal in preference for the typically disfavored 'P' and 'D' residues in the C-
terminal positions of the substrate.

20   The marked changes in preference of the D465A mutant toward the C-
terminal residues were not anticipated. However, it is known that the side chain
of D465 coordinates with ATP. Consequently truncating the side chain of D465
would be expected to perturb some aspect of ATP binding or function. No major
change in the Km for ATP, however, was revealed by analysis of the kinetic

25   parameters for D465A. Therefore, ATP contact with the remainder of the ATP
pocket within the enzyme may be sufficient for good binding in D465A.
However, the conformation of the enzyme's N-lobe may be abnormal due to a
lack of favorable interaction between the D465 side chain and other elements in
the N-lobe. This incomplete closure would be expected to alter the "closed

30   conformation" that the enzyme usually adopts during catalysis, and alter
movement of alphaC towards the activation loop.

**EXAMPLE 7: Analysis of different assay conditions with methods of the
invention**

Tests were performed on a wild type kinase to examine whether low ATP
concentrations would favor an ordered reaction in which a peptide binds first in

5   the absence of ATP, and subsequent loading of ATP rapidly proceeds to
catalysis. The PSSMLogo for such as assay is shown in FIG. 26. This
PSSMLogo for low ATP reveals a distortion of shape that bears substantial
resemblance to the D465A PSSMLogo. Specifically, there were decreases in
height of the P+2 and P+3 columns that are even more marked that those

10   observed with D465A. Moreover, like D465A, the low ATP profile has lost
many of the characteristic preferences of the other constructs at these positions
(see below).

Visualization of D465A preferences at individual positions was
facilitated by the graphical analysis shown in FIG. 27, which shows data for the

15   eight most informative residues at four particularly informative positions.
Positions P-2 and P-3 are shown in part because those are the peptide positions
at which the greatest changes resulting from point mutations of acidic residues
were anticipated. Positions P+2 and P+3 are shown because they are the location
of many of the biggest changes in D465A and low ATP conditions. The most

20   striking finding was the similarity in residue preference that occurs with D465A
and low ATP, but not for other mutants. There were fifteen such changes,
denoted with solid arrows below the x-axes in FIG. 27. Amongst these changes,
five occur in the N-terminal P-2 and P-3 positions. Two of these N-terminal
changes were ones that had been predicted, namely decreased preference for H at

25   P-3 and decreased disfavor for D at P-3. The failure to see decreased preference
for R or K at P-3 suggests that conformational flexibility allows binding of the
P-3 substrate residue to residues other than D465 in the cleft (most likely D544
or D508).

The correlation between the D465A and low ATP changes in the C-

30   terminal region of the substrate was striking. In almost all cases the changes in
substrate preference observed for D465A involve neutralization of the strong
preferences (either negative or positive) observed for related kinases. In
contrast to D465A, changes in substrate preference for the other three point
mutants are quite modest both in number and magnitude of change. However,

95

some changes in substrate preference for the D508A mutant bear similarity to those found in D544A (denoted with dashed arrows above the axes in FIG. 27). Both have lost their disfavor for D at the P-2 position (consistent with repulsion by nearby residues). Both also show a modest decrease in preference for R, not

5   only at P-2 but also at P-3.

The methods of the invention are therefore informative not only for studying the specificities of mutant kinase constructs, but also for analyzing changes in kinase specificity resulting from different assay conditions. It can be easily appreciated by one of skill in the art that the present methods would be

10   useful in analyzing importance of other assay conditions, such as ion
•
concentration (Ca++, Mg++, H+), and temperature. The present methods would also be useful in determining whether addition of other molecules to the assay influenced peptide specificity, for example by allosteric effects.

15   **EXAMPLE 8: Further understanding of anchor residues and their**
                                variations in test sets

Understanding of substrate specificity usually requires understanding the residue preferences at every position close to the phosphorylation position. The problem related to establishing anchor positions is that positions that are chosen

20   as anchor residues in a set cannot, by definition, also be query or variable positions in that set. For example, the peptide test set Rxx-S-F uses anchor residues at positions P-2 and P+1. Therefore, information on the P-2, P0 and P+1 positions cannot be obtained from the Rxx-S-F test set. In the embodiment shown in FIG. 2, the P-3, P0, and P+1 positions were analyzed by using

25   diminished numbers of anchor residues. For example, for the P+1 test set, the anchor at P-3 was retained, but the P+1 position was used as the query position (variable residue). Note that the methods of the invention provide strategies for designing and using a variety of test sets that could determine information about the residue preference for PKC-theta at the P+1 position. FIG. 28 illustrates

30   results with such varied test sets used for analysis of specificity of PKC-theta; each column of the PSSM logo represents results with a single test set and the symbolic representation of that set is shown below the column. Consider for example residue preference at the P+1 position, which our experience with the methods of the invention indicates is particularly important. Residue scores

determined for that position vary depending on the number (and position) of the anchor residues used in the test set. Also note that the results differ significantly for test sets in which the phosphorylatable residue is T rather than S. For one skilled in the art, the methods of the invention provide many strategies to refine

5   the definition of specificity for a kinase. For example, because the P+1 preferences for threonine phosphorylation differ from those for serine phosphorylation, one can create test sets analogous to those shown in FIG. 2, but using T as the phosphorylatable residue. Results with those peptides would allow more precise predictions, because they would be tailored specifically to

10 relevant subsets of peptide substrates.

FIG. 29 illustrates results with another superset of test sets of peptide pools based on a single anchor residue of R at P-3 and threonine as the phosphorylatable residue. Results shown are for the kinase ROK-alpha, about which there is little general understanding of specificity in the literature. This

15 superset is designed as a screening set to ascertain gross preferences from which to choose an additional anchor position. For that reason, it was most economical to only include 4 query residues: R, E, L and F, which our experience indicates are particularly important anchor residues . Even this limit analysis shows a strong overall preference for R, indicating ROK is clearly a "basophilic kinase".

20 The only position tested which has a dominant hydrophobic preference is P+3. One practiced in the art of this invention can appreciate that the third anchor position for a full test set of peptides should most likely be an 'R' at the P-4 or P-5 positions, where it has the strongest preference and where there are no other favorable residues.

25

EXAMPLE 9: Querying by Fixed Residue at Varied Positions rather than by Varied Residue at Fixed Position

The large family of basophilic kinases has a preference for arginine (R) at many positions in the substrate (see for example, FIG. 8, FIG. 13, FIG. 22,

30 and FIG. 29). Accordingly, arginine is a good candidate for an anchor residue at the high-scoring position(s). With this in mind, over-representation of arginine in anchor optimization sets used to assign anchor positions is a good first approach for an assay designed to assign anchor positions because the data

indicate that arginine can markedly enhance the efficiency of phosphorylation when it is present in a peptide substrate for such kinases.

In this Example, an anchor optimization set referred to as an "R-pair set" was created to systematically evaluate the use of arginine in each position

5   around P0 (in this set occupied by serine) from position P-7 to P+3. FIG. 30 shows the forty-five peptide sequences of this R-pair set. Results for the R-pair set using protein kinase A (PKA) are shown in FIG. 31. The results were calculated in a fashion similar to the sets described previously. Residue preference was calculated as follows:

10   [cpm for a peptide, calculated as the geometric mean for replicate values]/

                    [geometric mean cpm for all peptides in the set].
The position specific residue score was determined by calculating $\log_2$ of the residue preference. An average score for arginine at each position was also

15   calculated as the arithmetic average of the scores for all nine peptides that have a fixed arginine at the position. Inspection of the average score reveals that there PKA shows a strong overall preference for arginine at positions P-3 and P-2. Inspection of the results for individual peptides confirms that PKA most efficiently phosphorylates the individual degenerate peptide that has arginine

20   fixed at both P-3 and P-2. These results for PKA are in agreement with a summary of the literature, for example with results obtained by the Tegge approach to determining optimal kinase substrates (Tegge W et al. 1995. Biochemistry 34:10569-10577).

One simple way to summarize the results of studies with the R-pair set is

25   to determine the geometric average preference for all peptide pools that have R at a given position. For example, in this embodiment, there are 9 peptide pools that have R at P-3 (see FIG. 30 and FIG. 31). The geometric average preference for R in those 9 pools is 1.5 (FIG. 32). Similar calculations for the other positions, results in the graph shown for PKA in FIG. 32 which likewise

30   illustrates that PKA prefers R at P-2 and P-3.

Use of the R-pair set for anchor optimization with other kinases is likewise highly informative. For example, a comparison of the average position-specific scores for PKC-alpha and AKT1 with those described above for PKA is shown in FIG. 32. As shown in FIG. 32, PKC-alpha prefers arginine at P-3, P-2

and P+2. This is precisely the dominant positions at which the strongest

preference for basic residues have been found in a summary of literature results

for PKC (Kreegipuu A et al. 1998. FEBS Lett 430:45-50). Results from an R-

pair analysis with AKT1 show that arginine is preferably placed at positions P-3

5   and P-5 (FIG. 32); these results are in agreement with findings from the

literature (Obata T et al. 2000. J Biol Chem 275:36108-36115). Thus, the

strategy provided herein for efficiently scanning for critical residues provides

highly informative results. These residues are candidates for anchor residues for

more complete degenerate residue sets. One key advantage of this particular set

10   (and the approach of position scanning) is that it provides an impartial way to

assess the most important position for R without introducing biases from other

anchor residues. This general strategy of scanning for the optimal position of a

defined amino acid is referred to herein as "Optimal Residue Position Scanning"

(ORPS). The ORPS approach is further illustrated in Example 12 using arginine

15   and phenylalanine as the defined amino acids.

EXAMPLE 10: Detection of SHP-1 phosphorylation in whole cells

Prediction of phosphorylation sites is ultimately most useful to

understanding cellular physiology when it can be applied to facilitate

20   identification of sites that are relevant in intact cells. This Example illustrates

strategies for analyzing phosphorylation of the SHP-1 protein that extend the

information provided from the previously illustrated *in vitro* studies.

SHP-1 (also referred to as PTP1c, PTPN6 and SHPTP-1) is a tyrosine

phosphatase that critically regulates many signaling responses, including

25   activation of T-lymphocytes by the T-cell receptor (Okumura M et al. 1995.

Curr Opin Immunol 7:312-319; Kosugi A et al. 2001. Immunity 14:669-680).

The functioning of SHP-1, and especially its phosphatase activity, is modified by

phosphorylation. Sites thought to be phosphorylated include Y536 and Y564,

both of which are close to the C-terminus of the molecule (Zhang Z et al. 2003.

30   J Biol Chem 278:4668-4674).

SHP-1 has been shown to be a substrate for serine phosphorylation by

PKC (Zhao Z et al. 1994. Proc Natl Acad Sci U.S.A. 91:5007-5011).

Phosphorylation of SHP-1 by PKC results in decreased catalytic activity of SHP-

1 (Brumell JH et al. 1997. J Biol Chem 272:875-882). Other investigators have

shown that a closely related phosphatase, SHP-2, is phosphorylated on serine
residues close to its C-terminus (Strack V. et al. 2002. Biochemistry 41:603-
608). However, Strack et al. (*id.*) incorrectly inferred that SHP-1 is not
phosphorylated by PKC and previous studies have not identified the critical site
5   of phosphorylation by PKC.

Phosphorylation of SHP-1 was analyzed using the methods provided
herein, including the predictive algorithm for PKC-theta. Because
phosphorylation by PKC-theta correlates highly with that for PKC-alpha and
PKC-delta, these predictions have relevance at least for PKC-alpha and PKC-
10   delta, and likely provide a generalized prediction for novel and classical PKCs.

Table 7 provides the predictions made by the methods of the invention
for SHP-1 phosphorylation. For PKC phosphorylation using the fifth percentile
as a conservative cutoff that will include all plausible candidate sites for PKC
(see FIG. 9 and FIG. 11), only three sites in SHP-1 are predicted to be
15   phosphorylated (sites Ser-591 SEQ ID NO 298, Ser-26 SEQ ID NO 299 and Ser-
32, SEQ ID NO 300).

**TABLE 7. Three Predicted PKC Phosphorylation sites in SHP-1
whose corresponding phosphopeptides bind best to pPKC antibody**

Site					PKC-Theta	PKC-Zeta	PKA	pPKC antibody Score
Gene and Protein Name	SEQ ID NO	Phospho peptide Sequence		P0				
SHP-1	298	ADKEKSKG-pS-LKRK----		591	2	8	10	4
	299	LKGRGVHG-pS-FLARPSRK		26	0.3	0.8	10	2
	300	HGSFLARP-pS-RKNQGDFS		32	2	2	20	3
	289	MKNAHAKA-pS-RTSSKHKE		553	8	8	10	2
	290	RVILQGRD-pS-NIPGSDYI		294	60	60	10	2
	291	AHAKASRT-pS-SKHKEDVY		556	10	20	30	3
	292	KKKLEVLQ-pS-QKGQESEY		528	30	30	90	2
	293	PSEPGGVL-pS-FLDQINQR		431	50	50	30	2

294	HAKASRTS-pS-KHKEDVYE	557	8	7	2	1
295	PWTFLVRE-pS-LSQPGDFV	138	40	20	7	3
296	KNQGDFSL-pS-VRVGDQVT	42	10	20	50	3
297	PLNCSDPT-pS-ERWYHGHM	107	60	60	90	2

As shown in Table 3, a peptide that includes Ser-591 is phosphorylated by PKC (see SEQ ID NO:209, in Table 3). In particular the *in vitro* phosphorylation by PKC-theta was measured for the DKEKSKGSLKRK—(

5   SEQ ID NO:209) peptide and shown to be 17. A commercially available antibody from Cell Signaling Technology, referred to as a phospho-PKC motif antibody (designated herein as pPKC Ab), was used to generate the antibody binding data illustrated in Table 3. (See U.S. Patent 6,441,140 and Cell Signaling Technology Datasheet for 'Phospho-(Ser) PKC Substrate Antibody').

10   Information from Cell Signaling Technology indicates that this antibody preparation may recognize a motif consisting of positively charged residue at P-2, a serine at P0, a hydrophobic residue at P+1 and a positively charged residue at P+2. Such antibodies can be used for detection of unknown proteins that contain phosphorylation sites conforming to the motif to which they bind. For

15   example, phosphorylated proteins can be detected on two-dimensional gels with the pPKC Ab and the identity of these phosphorylated proteins can be confirmed by the observed molecular weight, isoelectric point and other information such as the predictive algorithms provided herein. Similarly, such detected proteins can be enriched by classical biochemical separations, and when sufficiently

20   enriched, can be identified by mass spectrometry (Astoul E et al. 2003. J Biol Chem 278:9267-9275).

One basis for predicting whether the pPKC antibody can bind to a particular phosphorylation site is the extent of its conformity with the motif described for the antibody: [RK]x-pS-[FYILMV][RK]. Therefore for each

25   candidate site in SHP-1, a score from 0 to 4 was calculated based on the number of matches of the sequence to that pattern. That "pPKC antibody score" is tabulated for pertinent SHP-1 sites in Table 7. Ser-591 is the only site in SHP-1 that has a perfect score of 4.

To test whether phosphorylation actually occurs at these sites in vivo, an antibody specific for the corresponding phosphorylated peptide can be used. However, because the identity of the relevant sites was previously unknown, no such specific antibodies were available. The inventor therefore devised an

5   alternative approach using the pPKC Ab. Although antibodies such as the pPKC Ab are poly-specific, they can be constrained to provide information on the phosphorylation state of a particular molecule such as SHP-1 by isolating the molecule of interest and then testing the antibody for reactivity with that isolated molecule. That strategy was implemented for SHP-1.

10   In particular, SHP-1 was immunoprecipitated from the cell lysate of the cell line JURKAT with an anti-SHP-1 antibody (C-19; from Santa Cruz Biotechnologies) and protein G beads. The purified SHP-1 was separated by standard polyacrylamide gel electrophoresis, transferred onto a membrane, and blotted with 2 different antibodies as shown in FIG. 15. Results from Western

15   blotting with the anti-SHP-1 antibody (C-19 from Santa Cruz Biotechnologies) demonstrate that SHP-1 was successfully isolated and that it had a molecular weight of 64kd, characteristic of SHP-1. That SHP-1 immunoprecipitate also reacted with the pPKC motif Ab, indicating that a phosphorylated site(s) exists on SHP-1 that conforms to the motif recognized by the pPKC antibody.

20   FIG. 15 also provides information on JURKAT cells stimulated to activate SHP-1 via a T-cell receptor. Specifically, Jurkat T Ag cells were stimulated with CD3 antibody (clone 38.1, IgM ascites, 1:1000 Final) plus CD28 antibody (clone 9.3, sup, 1:1000 final) for different times, as indicated in FIG. 15. The amount of phosphorylated SHP-1, detected by intensity of the band on

25   the pPKC antibody Western blot, increased markedly within the first minute following stimulation. These data demonstrate that the phosphorylation of SHP-1 at the sites recognized by the antibody is increased following T-cell receptor stimulation. Thus, the site(s) on SHP-1 detected by the pPKC antibody (Table 7) are biologically relevant for immune cell responses (FIG. 15).

30   Two lines of evidence strongly suggested that S591 was a functionally significant phosphorylation site on SHP-1: S591 was uniquely strong predicted to be phosphorylated by PKC, and S591 had a uniquely good fit to the pattern detected by the pPKC antibody.

To directly test the functional significance of S591, a SHP-1 construct was generated in which S591 was mutated to alanine (i.e. S591A mutation) to test whether SHP-1 was still phosphorylated in the absence of the S591 residue. The mutation was created using the Quikchange methodology from Stratagene.

5   Using similar methods, an A148E mutation was also made in PKC-theta to generate a construct encoding constitutively active PKC-theta. Wild type SHP-1 and S591A mutant SHP-1 were transfected into 293T cells using calcium phosphate transfection in the presence or absence of the constitutively active PKC-theta construct. The transfected cells were cultured for 24hr, lysed, and

10   analyzed by Western blot in a manner generally similar to FIG. 15. Two important results came from the analysis (FIG. 42). First, co-transfection of PKC-theta with wild type SHP-1 resulted in phosphorylation of SHP-1 as detected by the pPKC antibody. Second, such phosphorylation was absence in the S591A construct, indicating that S591 is a major, if not the major, site of

15   SHP-1 phosphorylation. These results therefore established that SHP-1 S591 can be phosphorylated by PKC-theta.

Although the pPKC antibody can identify important phosphorylation sites, the pPKC antibody is designed to recognize many different phosphorylation sites that have basic residues at P-2 and P+2. For example, as

20   described by its manufacturer, Cell Signaling Technology, the pPKC antibody binds to SEQ ID NO:229 (WKN-pS-IRH). Hence, the pPKC antibody is not particularly site-specific.

Therefore a site-specific phospho-antibody was generated. A phospho-peptide having sequence CDKEKSKG-(pS)-LKRK-OH (SEQ ID NO:570) was

25   made. This phospho-peptide includes a sequence that corresponds to the C-terminus of SHP-1 but, in addition, it has an N-terminal cysteine useful for coupling to a carrier. The corresponding non-phosphorylated peptide was also synthesized for use as a control. The phospho-peptide (SEQ ID NO:570) was coupled onto the carrier KLH, rabbits were immunized, and anti-sera samples

30   were screened for reactivity with the phospho-peptide by ELISA assay. Antibodies reactive with corresponding non-phosphorylated peptide were removed from anti-sera by passing the anti-sera through a column having the non-phosphorylated peptide bound to the column matrix. Finally, anti-sera were

enriched for phospho-specific reactivity by use of an affinity column made from the phospho-peptide.

The specificity of the antibody for SHP-1 pS591 was confirmed by Western blot analysis (FIG. 43). When the anti-SHP-1 pS591 antibody was used

5 at a dilution of 1:15,000, only a single strong band was detected on a Western blot of a lysate of Jurkat cells. The position of this band was characteristic of SHP-1. In contrast, the pPKC antibody bound to many bands. Binding of the anti-SHP-1 pS591 phospho-antibody depended entirely on S591 because no such binding was detected in lysates of cells that expressed the SHP-1 S591A mutant

10 (co-transfected with constitutively-active PKC-theta). Thus, unlike the pPKC antibody, this anti-pS591 antibody had narrow specificity and was sufficiently specific for detection of only SHP-1 S591 phosphorylation. Prior immunoprecipitation of SHP-1 was not needed when the anti-pS591 antibody was employed. The strong reactivity of this antibody with phosphorylated SHP-

15 1 facilitated demonstration that CD3 cross-linking stimulates phosphorylation of SHP-1 both in the cultured cell line JURKAT cells and in normal mouse thymocytes.

PKC inhibitors were then used to further confirm that PKC mediates CD3/2-induced phosphorylation of SHP-1 (FIG. 44). Jurkat cells were

20 stimulated with CD3/CD28 after pre-treatment with graded concentrations of two PKC inhibitors: BIM I and BIM III. As shown in FIG. 44, SHP-1 phosphorylation was reduced by 1 micromolar concentrations of BIM I and BIM III and was virtually abolished at BIM I and BIM III concentrations of 5 micromolar.

25 The specificity of the anti-SHP-1 pS591 antibody was also demonstrated by *in situ* immunofluorescence studies (FIG. 45). Experiments were conducted with a wildtype and S591A constructs of SHP-1 N-terminally tagged with the fluorescent marker GFP. These constructs were transfected into 293T cells, the cells were then cultured for 24hr, fixed, permeabilized, and stained.

30 Immunofluorescent staining for SHP-1 phosphorylation was performed by incubating cells first with rabbit anti-pS591 and subsequently with an anti-rabbit antibody linked to the Alexa 568 fluorophore. FIG. 45 shows staining by anti-pS591 antibodies of cells transfected with wild type SHP-1 but not of cells transfected with S591A SHP-1.

Further investigation of the subcellular localization of SHP-1 in Jurkat
cells indicates that phosphorylation regulates the ability of SHP-1 to translocate
into the nucleus. FIG. 46 illustrates that C-terminally GFP-tagged SHP-1 (seen
as a light stain, green in the original) was located primarily in the nucleus. The

5   S591A mutant of SHP-1 was also detected in the nucleus, but the S591D mutant
was largely excluded from the nucleus. The change in SHP-1 of S591 to D591
mimics phosphorylation at residue 591, and caused exclusion from the nucleus.
Moreover, in 293T cells co-transfected with SHP-1 and constitutively active
PKC-theta (which causes phosphorylation of SHP-1 S591, see FIG. 43) results

10   in exclusion of SHP-1 from the nucleus. However, incubation of SHP-1/PKC-
theta expressing cells with the PKC inhibitor BIM I causes the SHP-1 to become
localized within the nuclei (FIG. 46B). Also, as shown in FIG. 46C, the ability
of PKC-theta to cause exclusion of SHP-1 from the nucleus is destroyed by
mutation of S591 to alanine (A). Thus, multiple lines of evidence indicated that

15   phosphorylation of S591 cause exclusion of SHP-1 from the nucleus.

**EXAMPLE 11: Additional examples of proteins predicted to have good
PKC phosphorylation sites and found to bind pPKC antibody by Western
blot**

20   The predictive power of the methods of the invention is further illustrated
in this Example by studies of the proteins LIMK-2 and MLK3. LIMK-2 and
MLK3 were identified as promising candidates for phosphorylation by PKC
based on predictions for PKC-theta described herein and confirmation of that
prediction by *in vitro* peptide phosphorylation (SEQ ID NO: 76 in Table 4 and

25   SEQ ID NO: 121 in Table 5).

In vitro binding experiments were performed to determine whether the
pPKC Ab bound to predicted phosphorylated sites in MLK3 and LIMK2.
Synthetic peptides chosen from those shown in Table 4 were subjected to
phosphorylation by PKC-theta. Assay conditions were similar to those

30   described herein, except that the phosphorylation reaction was for 30 minutes at
30 °C and then overnight at 4 °C. The reaction mixture was applied to HB
avidin-coated plates, the plates washed, and then pPKC Ab binding was
determined. The results of these assays are summarized in Table 8.

TABLE 8. The pPKC Antibody binds to peptides after phosphorylation by PKC-theta

Gene name	SEQ ID NO	Sequence	pPKC Ab Signal			Peptide phosphorylation by PKC-theta
			on peptide without exposure to PKC-theta	on peptide after exposure to PKC-theta phosphorylation	amount dependent on PKC phosphorylation	
MLK 3	76	HVRRRRGT FKRSKLRA RD	0.07	1.02	0.95	99
LIMK -2	121	LRRRSLRRS NSISKSPGP	0.02	1.13	1.11	57
ROC K2	75	EEAEHKAT KARLADK	0.02	0.02	0.00	0

As shown in Table 8, the pPKC Ab bound to peptides from LIMK-2 and from MLK3 after phosphorylation but not before. Results for a control peptide (ROC K2) are also shown; the ROCK2 peptide is not phosphorylated by PKC and shows no change in binding to pPKC Ab after the peptide was exposed to PKC-theta.

The question of in vivo relevance of LIMK-2 phosphorylation was addressed using the strategy used above for SHP-1. LIMK-2 was immunoprecipitated with anti-LIMK2 antibody H-78 purchased from Santa Cruz Biotechnologies, separated by one-dimensional PAGE and analyzed by Western blot. The Western blot shown in FIG. 33 illustrates that LIMK-2 was immunoprecipitated from T-lymphocytes before and after T-cell receptor stimulation and the pPKC antibody bound to LIMK-2, indicating phosphorylation of LIMK-2. Note that the pPKC signal was observed only on the sample from T-cell receptor stimulated cells, indicating that phosphorylation of LIMK-2, as detected by the pPKC antibody, occurred during T-cell receptor stimulation.

Similar studies were performed with the MLK3 protein. Jurkat T Ag cells (10 million) were stimulated with CD3 (clone 38.1, IgM ascites, 1:1000 Final) plus CD28 (clone 9.3, sup, 1:1000 final), or with PMA (200ng/ml) for 5 minutes. MLK3 was immunoprecipitated from the cell lysate with anti-MLK3 Ab (H-300; from Santa Cruz) and protein G beads. The immunoprecipitated MLK3 was

subjected to western blotting and one blot was probed with the pPKC Ab while another blot was probed with the MLK3 Ab. As shown in FIG. 34, MLK3 has strong reactivity with the pPKC antibody both before and after stimulation of JURKAT cells. The predicted phosphorylation site at Ser-477 on MLK3

5   corresponds to one of the very best detected in the entire human proteome, and the JURKAT cell line is a partially activated transformed cell line. The binding of pPKC antibody therefore likely reflects phosphorylation of MLK3 that is present even in unstimulated cells.

10   **EXAMPLE 12: Evaluation of best positions for arginine and phenylalanine in an RF-pair peptide set for PKC-theta phosphorylation**

Example 9 introduced the idea of "Optimal Residue Position Scanning" (ORPS) using pairs of R residues at all possible positions near P0. This Example further illustrates the ORPS approach including the design, synthesis

15   and testing of a set of degenerate peptides in which a single arginine and a single hydrophobic (phenylalanine) residue are the only two fixed residues near a phosphorylatable residue (S at P0). Arginine was chosen for this analysis because of its importance to basophilic kinases. A hydrophobic residue was chosen as the second residue because a synthesis of the scientific literature

20   indicated that one or a few hydrophobic residues are often important determinants of the specificity of multiple kinases. For example, several PKCs have an apparent preference for a hydrophobic residue at P+1. While a variety of hydrophobic residues exist, including, for example, phenylalanine or leucine or a mixture of several residues (such as isoleucine, leucine, methionine, valine

25   and/or phenylalanine), for this proof of principle a single hydrophobic residue (F) was selected to maximize informative design consistency between this set and the RxxSF set.

Design details for the RF-pair set are illustrated in FIG 36. As in other peptide sets, each peptide consisted of an N-terminal linker (biotin-dansylated

30   lysine and glycine) followed by a 13 residue insert. The insert consisted of a fixed serine residue flanked by eight N-terminal residues and four C-terminal residues. Each peptide had a single R at a position ranging between P-7 to P+4 and a single F at another position ranging between P-7 and P+3. The symbolic representation of two such peptides is shown in FIG 36. Altogether the peptide

set included all possible combinations of R and F at positions between P-7 to P+3 (excluding P0).

The specificity of PKC theta for various peptides was assessed using PKC-theta phosphorylation reactions with peptides of the set then calculating log

5   scores as described above for the R-pair set. In FIG. 36, scores showing distinctly favored phosphorylation (>0.5) are highlighted with bold and underlined while those showing distinctly disfavored phosphorylation (<-0.5) are bold but not underlined. Visual inspection of the results indicates underlying patterns. The position most favored for R is P-2 because 7 of 9 peptides in that

10 column are distinctly favored. The P-3 position is also favored for R (4 of 9 peptides distinctly favored). The position P+1 is clearly most favored for F because 8 of 10 peptides in that row are distinctly favored.

An alternate way to assess residue preference at a position is by determining the average score for all peptides sharing that residue at that

15 position. Those values are shown in the right hand column and the bottom row of FIG. 36. FIG. 37A provides a graph of the average position-specific preferences of PKC-theta. As shown in FIG. 37, analysis of the RF pair set indicates that P-2 is the preferred position for R and P+1 the preferred position for F. These results for arginine are similar to those obtained in Example 9 for

20 arginine alone. Thus, analysis of PKC-theta with the R-pair set (FIG. 37B) also indicates that the P-2 is the single most important position for an R residue in PKC substrates.

As indicated in previous Examples, analysis of PKC-theta with the RxxSF set of peptides was quite informative. It seems likely that analysis of

25 peptide specificity will be even more informative when "systematic amino acid variation on template substrate" (SAaVoTS) is used to design better peptide sets (e.g. RxSF). Thus, the R-pair and RF-pair sets serve the critical purpose of objectively determining what are good residue choices for positional scanning approaches (SAaVoTS). (See also Example 14).

30 FIG 38 shows the distribution of log2Scores for the PKC-theta with the RF-pair set, sorted from highest to lowest scores. As shown in FIG. 38, there are 4-7 peptides that are distinctly superior in their phosphorylation, rather than a single peptide in the RF-pair set that is exceptionally well phosphorylated. This is consistent with complex additive or alternative modes of binding of substrate.

If particularly high resolution analysis of specificity of PKC-theta is required, then analysis with SAaVoTS sets based on several of these RF-pair peptides is likely to provide additional information.

5   **EXAMPLE 13: Analysis of kinases with a "diverse basic proteomic set," which is enriched in for sequences located near the N- and C-termini of proteins.**

Although degenerate peptides are particularly useful for studying kinase peptide specificity, strategic use of non-degenerate peptides can also be

10   effective. Thus, a set of 96 peptides with defined sequences was designed and synthesized, each comprised of a preferred N-terminal linker and a 17 residue insert (Table 9). The inserts were chosen by the following criteria. First, only sequences from human proteome were selected. Second, peptide choice was biased towards sequences that basophilic kinases favor for phosphorylation,

15   especially PKC-theta, using the prediction methods described herein. Consequently the sequences were enriched in basic residues: R was enriched in the peptides to an abundance of 19.3%, more than three-fold higher than that observed in the human proteome (about 6%); and K was enriched in the peptides of the set to 12.3%, more than two-fold higher than observed in the human

20   proteome. Moreover, 80% of the peptides were in the top 5 percentiles for predicted phosphorylation by PKC-theta. Third, the diversity of the peptides was enhanced by manually selecting sequences having diverse residues at positions strongly biased by the PKC preference (especially diversity at the P-2, P-3, P-4) positions. Fourth, the set was enriched for peptides corresponding to

25   proteins that are well expressed in hematopoietic cells so that findings would be most relevant to the inventor's field of interest. Fifth, the peptide set was enriched for sequences at or near the C-terminus of the protein (46 of the 96 peptides) and the N-terminus (5 of 96 peptides). This choice to emphasize C- and N-terminal peptides was made based on the knowledge that sequences near

30   the termini of proteins are the most mostly likely to be available for interactions with other proteins. Although the accessibility of protein termini is best know in the context of protein immunization/detection with antibody, the data illustrated herein indicate that the same principle applies to the accessibility of termini for interactions with other proteins (such as kinases). Moreover, there is

experimental evidence that basic residues at the C-termini have special

importance to protein function (Scheglmann D, Werner K, Eiselt G, Klinger R.

2002. Protein Eng 15:521-528). Sixth, the peptide set was enriched for

sequences that were not strongly hydrophobic (see Table 9 column

5   "hydrophobic; the hydrophobicity scores for individual residues are shown in

FIG. 14). The mean hydrophobicity of peptide sequences from the human

proteome that have 17 residues is about 0.34, while the mean hydrophobicity of

the 96 peptides in Table 9 was in the fifth percentile for the proteome ($\leq$ -0.07).

The selection of hydrophilic peptides further enhanced the likelihood that these

10   sites would be accessible for phosphorylation and functional interaction in native

proteins.

Table 9

Sequence	Seq ID No	Locus LinkID	Name	Possible P0	N- or C-Term	Hydro-phob-icity	AKT1	PKC-theta	PKC-zeta	PAK	MST4
Novel PKC phosphorylation sites in peptides with sequence found near N- or C-termini of intracellular proteins										
GTIRSKLSRRRS AQMRV	474	547	ATSV/KIF1A	1685	C-term	-0.14	3	**60**	**23**	8	**100**
SPGRRRHRSRSS SNSSR	475	10921	RNPS1	299	C-term	-0.45	**95**	**57**	7	**20**	2
TSGVRRRRSQRI SQRIT	476	25836	IDN3	2150	C-term	-0.08	5	**83**	**59**	8	1
PQKSSFFSLRRK SRSKD	477	1608	dag kinase gamma	787	C-term	-0.05	10	**80**	**24**	**31**	10
SSLAQRRSMKK RTSTGL	478	27330	p90-RSK6	741	C-term	-0.14	4	**68**	**38**	10	2
KYSVKRKKSRS KKVRRH	479	9014	TAF1B	580	C-term	-0.39	5	**83**	11	**11**	**11**
GREREREKRFSF FKKNK	480	6712	spectrin, beta, non-erythrocytic 2	2384	C-term	-0.28	5	**56**	**19**	**13**	1
RERRRNERLRRE SVRPV	481	941	CD80	284	C-term	-0.16	7	**44**	**26**	9	2
FFKRRKRKSLQR HK	482	1455	casein kinase I gamma 2	410	C-term	-0.2	4	**45**	8	9	1
VLRTRHARHTR KRRRLV	483	6621	SNAPC4	1462	C-term	-0.15	5	**47**	3	5	2
KRRGGRRRSRSP DRRRR	484	9939	RBM8A	168	C-term	-0.61	**12**	**29**	2	**12**	1
QKPVMVKRKRT RPTKSS	485	6158	RPL28	131	C-term	-0.03	4	**26**	3	6	1
ESDHQIIKRRLRT KTAK	486	9585	MPP1	1776	C-term	-0.01	10	**16**	2	5	3
NKRLREKRVSNS KFYS	487	5336	PLCG2	1259	C-term	-0.11	3	**15**	7	8	1
RKSNLERHHRL HTGKKS	488	55762	FLJ10891	461	C-term	-0.05	3	**20**	3	2	1
IKPRNITRRKTD REEKT	489	2889	RAS-GRF2	1071	C-term	-0.36	2	**13**	2	3	1
EGDRLGRRSSSK RALKA	490	117532	TMC2	19	N-term	-0.21	4	**55**	**18**	6	1
NNHMKTKASVR KSFSED	491	11215	AKAP220	14	N-term	-0.1	5	**21**	9	**39**	1
RIIRPRPPSRSRA AQSP	492	22899	ARHGEF15	26	N-term	0.1	3	**36**	2	7	0
Other novel PKC phosphorylation sites in peptides with sequence from intracellular proteins										
LDNLKRKNTRR SIKLDG	493	10788	IQGAP2	1458		-0.1	5	**82**	**100**	7	4
DPRWRKRMSLT LKSNKN	494	1620	DBCCR1	525		-0.01	**13**	**37**	**15**	**100**	**23**
GDDFLRRSSSRR NRSIS	495	9595	PSCDBP	298		-0.12	9	**81**	**65**	**17**	3
TSRATRRKTNRS SVKTP	496	9656	NFBD1	1642		-0.2	6	**100**	**37**	7	9
YHNERARKSRN	497	785	CACNB4	470		-0.23	**16**	**28**	**15**	**69**	1

RLSSSS											
GRAYRHRGSLV NHRHSH	498	9726	KIAA0296 gene product	479		0	**14**	**61**	**22**	**14**	3
SHQGRNRRTVT LRRQPV	499	54221	SNTG2	73		-0.04	5	**57**	9	4	**34**
ASRRPPRRSRSR RPGLH	500	22947	DUX4	822		-0.17	**13**	**69**	**11**	9	1
RKARLSRRSRSA SSSPE	501	23524	SRm300	1707		-0.23	7	**49**	**28**	**16**	1
LHLPRGRRSRRA VSVPA	502	8471	IRS4	422		0.17	4	**44**	**17**	**30**	1
LDSGRKTRSARR RTTQI	503	4926	NUMA	1792		-0.15	10	**55**	**15**	4	9
SAQERLTRTFTR SSHTY	504	2318	gamma-filamin	2157		0.11	**21**	**38**	**15**	4	2
TTRGRKNRSSVK TPETV	505	9656	NFBD1	1522		-0.13	4	**45**	7	**12**	9
KSGTKVSRTFSY IKNKM	506	11214	AKAP13	1748		0.11	4	**41**	**21**	5	4
PSIRPRPGSLRSK PEPP	507	9744	centaurin beta 1	554		0.18	3	**49**	**19**	3	1
AERSRHQRSFSV PKKFG	508	10129	Hypothetical protein CG003	1980		0	10	**33**	**19**	7	1
EPNRIPSRSLRRT KLNQ	509	9656	NFBD1	1880		-0.02	4	**45**	**15**	4	3
KPKIRTRKSSRM TPFPA	510	9656	NFBD1	1112		0.1	6	**46**	**13**	3	1
RKNSARKASIVK NLKDT	511	4690	NCK1	66		-0.14	6	**40**	8	4	1
DMVEKTRRSLT VLRRAQ	512	862	CBFA2T1	347		0.14	3	**37**	7	8	1
YEGWYRGYTLR KKSKKG	513	1793	DOCK1	50		0.03	4	**38**	5	3	0
PREMKGKKSKKI SLKYT	514	8826	IQGAP1	1556		-0.06	3	**30**	3	5	2
TTAQPTKKSTLK KRVAR	515	926	CD8beta	148		0.02	3	**27**	2	3	5
NMLSLRHNSRIH VRPSR	516	10198	MPHOSPH 9	453		0.19	6	**19**	4	2	1
QVRTRIKRSRKL ADSNF	517	8842	CD133	190		-0.04	4	**14**	2	3	1
PKC phosphorylation of peptides from extracellular proteins											
FSNRGSVRTRRF SKHLL	518	2799	glucosamine (N-acetyl)-6-sulfatase	544	C-term	0.17	**14**	**95**	**93**	**20**	**25**
PKC phosphorylation of peptides corresponding to previously reported PKC sites found near N- or C-termini of intracellular proteins											
LSGLTWQRRQR KSRRTI	519	3559	CD25		C-term	-0.15	10	**85**	**29**	4	5
INKKAYEMASH LRRSQY	520	5216	Profiling I	138	C-term	0.12	2	**30**	**18**	7	0
RRLKKGKKSKR SMNDPY	521	6402	CD62L	364	C-term	-0.28	3	**26**	8	9	1
YEMLAARKKKV SSTKRH	522	9267	cytohesin-1	392	C-term	-0.15	2	**32**	2	3	1
PKC phosphorylation of peptide corresponding to previously reported phosphorylation site for another kinase											

DNKLRRYTTFSK RKTGI	523	6722	SRF	160		-0.02	**53**	**33**	**17**	4	**21**
				peptides not well phosphorylated by PKC							
ILLKRSGKSLNK EWKKK	524	116988	CENTG3	381		0.05	5	10	3	3	1
RRPREKRRSTGV SFWTQ	525	4659	MYPT1	852		-0.03	**12**	10	2	**16**	0
RRGLKRSLSEME IGMVV	526	6722	SRF	103		0.28	3	10	5	4	0
PSQKKKKKKKK TAEQTV	527	6949	TCOF1	1406	C-term	-0.48	3	10	5	5	3
QARQSRRSTQG VTLTDL	528	4659	MYPT1	696		0.07	4	10	4	7	0
RKPDRRKRSRPY KAKRQ	529	8314	BAP1	721	C-term	-0.49	3	9	2	4	1
TEVRERRRSYLT PVRDE	530	4659	MYPT1	668		-0.08	3	8	3	**17**	1
SRLGRGKRSLVL DLKQP	531	23600	AMACR	52		0.16	3	7	6	9	2
EQNVPKRRSRNT AVEQR	532	6840	Supervillin	324		-0.25	4	7	1	4	1
ARNESRSGSNRR ERGAP	533	7456	WIP	488	C-term	-0.37	2	7	2	2	0
LTEKGKHGSFLV RESQS	534	5781	SHP-2	134		0.13	2	7	2	3	1
NYGKKKKRRSRE KHQEST	535	6932	TCF1	234		-0.46	3	6	2	4	1
EKFSKTKKSKRK LEVDS	536	5978	REST	519		-0.24	2	6	2	3	1
RERRAPRRTRVN GDNRL	537	5335	PLC-gamma 1	1282	C-term	-0.32	4	6	1	6	1
EGIYRKSGSTNK IKELR	538	4649	Myosin IXA	2102		-0.03	3	6	2	3	1
AARARRIRRRTD VRITG	539	51282	SCAND1	173	C-term	-0.15	6	6	2	5	1
PMTDKRVASLL KKAKAQ	540	4297	MLL	1025		0.1	2	5	2	2	1
QPRKKRLKSIEE RQLLK	541	8805	TIF1	1007	C-term	-0.07	**12**	5	4	10	1
LLWWRRRKGSR APSSDF	542	29990	Paired immunoglobulin-like receptor beta	220	C-term	0.19	3	5	2	5	0
DRSYSHHRSPSE SSRYS	543	4820	NKTR	1456	C-term	0	4	4	2	6	1
PAFGPRRGSSPR GAAGA	544	7225	TRPC6	13	N-term	0.12	4	4	2	3	0
EDGDDLLHHHH VSGSRR	545	4000	lamin A	568	C-term	0.17	3	4	1	3	0
SSILAQRRVRKL PSTTL	546	6195	p90-RSK1	732	C-term	0.13	1	4	3	4	2
AQEPPAKRRKGS KGPEG	547	4968	OGG1	340	C-term	-0.23	3	4	5	6	1
QRSQLVVHRRT HTGEKP	548	27300	zinc finger protein	711	C-term	0.15	2	3	1	4	0
AKEMLTKKSTL RDQINS	549	10946	SF3A3	31		0.06	4	3	2	4	2
SAFPLSRKNKGS	550	28511	I-kappa-B-	186	C-term	-0.03	2	3	4	2	1

GSLDG			interacting Ras-like protein 2								
SRSYLIAHQRSH TREKL	551	10308	ZNF267	739	C-term	0.16	4	3	1	4	1
ELVASGHKKET QKGKRK	552	9603	NRF3	688	C-term	-0.33	3	3	3	7	1
QGRRKGRITRSM TNEAA	553	9611	NCOR	586		-0.14	5	2	1	6	0
LQQEPRKISYSRI PESE	554	6654	SOS1	1064		0.13	2	2	3	3	0
MSGRPRTTSFAE SAKPV	555	2932	GSK-3beta	9	N-term	0.25	_84_	2	1	7	0
INARRNRITLTL DDEAA	556	129684	caspr5 protein	463		0.12	3	2	2	3	0
ERRRRNKMTAY ITELSD	557	405	ARNT	106		-0.07	2	2	2	2	1
STLAQRRGIKKI TSTAL	558	6197	p90-RSK2	736	C-term	0.14	2	2	3	4	2
SGGKEEKKHHK SSDKHR	559	57187	Tho2	1509	C-term	-0.55	2	2	1	4	1
KKRRRQRQSGV VVEEPP	560	51755	CRK7	1053		-0.16	3	2	3	_33_	0
TLFQFSDSELRH GRSDQ	561	9252	MSK1	807	C-term	-0.24	2	2	1	2	0
EASARSPRSYLV AGANP	562	5241	Progesterone receptor	403		0.19	2	2	1	2	0
GKRYKFVATGH GKYEKV	563	4046	LSP1	323	C-term	0.07	5	1	5	5	0
KMGVAAHKKSH EESHKE	564	6279	MRP-8	86	C-term	-0.09	4	1	2	4	0
EQGNIVTKKDHT STPNP	565	286	Ankyrin R	104	C-term	0.1	3	1	1	4	0
QPRRKRLKSDER PVHIK	566	51592	TRIM33	1102	C-term	-0.14	_14_	1	1	5	1
LHVLFKKRTKSK EDEEG	567	8748	ADAM20	720	C-term	-0.18	1	1	2	4	0
GAYFDKAKEKS PIQSQI	568	6332	SCN7A	1676	C-term	0.17	2	1	1	4	0
SDDPEKLSRRSH DLHTL	569	9744	centaurin beta 1	734	C-term	0.07	4	1	1	3	0

This set of peptides is very useful for identifying new sites for basophilic kinases because the set has many potential phosphorylation sites (total=310) and the set diversely represents many patterns of residues, including basic residues around the phosphorylation sites. Seventy-six of its peptides (79%) include 2 arginines within 6 positions of a Ser/Thr, and 56 of its peptides (58%) include 3 arginines within 6 positions of a Ser/Thr. This is much higher than the frequency of these patterns in 17-residue peptides in the human proteome, which is 18% for 2 arginines (4 fold lower than in this set) and 5% for 3 arginines (12 fold lower). Thus, the probability of assembling a peptide set with 4 fold higher abundance of

this pattern by chance alone is vanishingly small, even for a set of only 10 peptides, much less a set of 96. Hence, the usefulness of this set is related to the purposeful enrichment of arginines in diverse positions near the Ser/Th phosphorylation site.

5   Table 9 also tabulates results from phosphorylating this panel of peptides with 5 different kinases. Phosphorylation results for each peptide are expressed as percentage of phosphorylation of the best substrate by the same kinase. The kinases AKT1, PAK1 and MST4 were purchased from Cell Signaling Technology and assayed according to the protocol provided by the manufacturer

10   ProQinase.

Table 9 illustrates that a high frequency of peptides are phosphorylated by PKC-theta (50 out of 96) and to a lesser extent PKC-zeta (27 out of 96). These results are not surprising based on the selection of peptides with sites having scores in the top 5 percentile for PKC-theta.

15   One useful finding was that many peptides (i.e. more than ten) were phosphorylated by two basophilic kinases AKT1 and PAK1, even though the peptides in this set were not specifically selected to provide substrates for those kinases. Thus, the intentional selection of a diverse distribution of arginines around the phosphorylation site provided an enriched set of peptides that

20   effectively acted as substrates for these kinases. For example, AKT1 phosphorylated 13/96 peptides but only one peptide (from GSK-3) was intentionally chosen as a control for AKT1 phosphorylation. Similarly, PAK1 phosphorylated 16/96 peptides.

Of particular note, six peptides were substrates for the kinase MST4,

25   which was previously not known to be basophilic. Ongoing analysis using the approaches described herein indicates that MST4 is basophilic and prefers basic residues at positions P+4 to P+6 (data not shown). These newly identified peptide substrates are useful for development of better *in vitro* kinase assays. This is particularly true for MST4, because a good peptide substrate has not yet

30   been identified for MST4.

Importantly, the peptide set of Table 9 constitutes likely candidates for in vivo phosphorylation in native proteins in vivo because these sites are located near protein termini.

This "diverse basic proteomic set" can also be useful in analysis of residue preference of basophilic kinases, as included in Example 14 below

EXAMPLE 14: Analysis of a kinase whose specificity is poorly defined
5   **with the RF-pair, the R-pair and the diverse basophilic proteomic set.**

This Example illustrates the specificity of PAK1, as proof of principle that the inventive methods enable better characterization of a basophilic kinase whose specificity was previously incompletely defined. PAK1 belongs to the STE20 family of Ser/Thr kinase.

10   FIG. 39 shows the analysis of PAK1 with the R-pair set. These results illustrate the singular and consistent importance of R at the P-2 position to PAK phosphorylation.

FIG. 40 shows analysis of PAK1 with the RF-pair set. The analysis of average preference from this set also strongly affirms the singular importance of

15   R at P-2; and also indicates a modest average preference for F at P-1, P+1 and P+3. Looking at the results for individual peptides in the set, it is apparent that each of the peptides RRxS, RRS, RFS, RRxxS and RxSxxF are each strongly favored. Thus, each of these peptide sets could be used as the basic for a SAaVoTS degenerate set for more detailed analysis of PAK specificity.

20   Analysis of PAK with the "diverse basic proteomic set" proved to be informative. Table 9 includes a tabulation of the results of phosphorylation of that peptide set by PAK. Underlying sequences patterns were analyzed to differentiate between substrates motifs that are phosphorylated (i.e. >10% of the best substrate) and those that are poorly phosphorylated ($\leq$ 10% of the best

25   substrate). The most informative results for PAK demonstrate that R at position P-2 is singularly important for phosphorylation of peptides in this set by PAK (FIG. 41). FIG. 41A shows the procedure for a chi-square analysis to determine whether arginine at position P-3 (relative to a phosphorylation site) contributes to phosphorylation of the 16 positively phosphorylated peptides. FIG. 41A

30   tabulates the results: 10 of the phosphorylated peptides have arginine at position P-3 while 6 do not; 45 of the non-phosphorylated peptides have arginine at position P-3 and 35 do not. The bottom half of FIG. 41A shows the calculation of expected distribution of peptides if the R at P-3 and the phosphorylation are independent of each other. The bottom row tabulates the probability (from a chi-

square test) that the R at P-3 is correlated with phosphorylation. In the case of R

at P-3, there is no significance to the correlation (p~0.6). In the case of R at P-2,

the probability is very significant (p<0.0001). All of the 16 phosphorylated

peptides comprise a site with R at P-2 relative to an S or T (shown in FIG. 41B);

5   in contrast less than half of the non-phosphorylated peptides have that pattern.

FIG. 41C shows the p-values for analysis of R at all positions between P-6 and

P+3; the results demonstrate that R at P-2 is unique in its importance.

Thus the R-pair analysis, the RF-pair analysis and analysis with the

"diverse basic proteomic set" each show that the P-2 position occupies a place of

10   dominant importance in determining kinase specificity. The consistency

between these independent approaches is strong evidence for their validity as

well as for the validity of the finding that R at P-2 is unusually important to

PAK. It is notable that the approaches provided herein provide more precisely

define kinase sequence specificity regarding this most critical location of basic

15   residues than is provided by previous workers (Tuazon PT, Spanos WC, Gump

EL, Monnig CA, Traugh JA. 1997. Biochemistry 36:16059-16064).

All patents and publications referenced or mentioned herein are

indicative of the levels of skill of those skilled in the art to which the invention

20   pertains, and each such referenced patent or publication is hereby incorporated

by reference to the same extent as if it had been incorporated by reference in its

entirety individually or set forth herein in its entirety. Applicants reserve the

right to physically incorporate into this specification any and all materials and

information from any such cited patents or publications.

25   The specific methods and compositions described herein are

representative of preferred embodiments and are exemplary and not intended as

limitations on the scope of the invention. Other objects, aspects, and

embodiments will occur to those skilled in the art upon consideration of this

specification, and are encompassed within the spirit of the invention as defined

30   by the scope of the claims. It will be readily apparent to one skilled in the art

that varying substitutions and modifications may be made to the invention

disclosed herein without departing from the scope and spirit of the invention.

The invention illustratively described herein suitably may be practiced in the

absence of any element or elements, or limitation or limitations, which is not

specifically disclosed herein as essential. The methods and processes illustratively described herein suitably may be practiced in differing orders of steps, and that they are not necessarily restricted to the orders of steps indicated herein or in the claims. As used herein and in the appended claims, the singular

5   forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "an antibody" includes a plurality (for example, a solution of antibodies or a series of antibody preparations) of such antibodies, and so forth. Under no circumstances may the patent be interpreted to be limited to the specific examples or embodiments or

10   methods specifically disclosed herein. Under no circumstances may the patent be interpreted to be limited by any statement made by any Examiner or any other official or employee of the Patent and Trademark Office unless such statement is specifically and without qualification or reservation expressly adopted in a responsive writing by Applicants.

15   The terms and expressions that have been employed are used as terms of description and not of limitation, and there is no intent in the use of such terms and expressions to exclude any equivalent of the features shown and described or portions thereof, but it is recognized that various modifications are possible within the scope of the invention as claimed. Thus, it will be understood that

20   although the present invention has been specifically disclosed by preferred embodiments and optional features, modification and variation of the concepts herein disclosed may be resorted to by those skilled in the art, and that such modifications and variations are considered to be within the scope of this invention as defined by the appended claims.

25   The invention has been described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form part of the invention. This includes the generic description of the invention with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is

30   specifically recited herein.

WHAT IS CLAIMED:

1.   A test set for characterizing substrate specificities of kinases comprising at least two peptide pools, wherein substantially every peptide in each of the peptide pools comprises one phosphorylatable amino acid position,

5   one query amino acid position, and at least one degenerate amino acid position, and wherein:

            (a) each peptide of every peptide pool has an identical phosphorylatable amino acid that can be phosphorylated by a kinase at the phosphorylatable amino acid position;

10   (b) the query amino acid position is at a defined position relative to the phosphorylatable amino acid position within every peptide of every peptide pool but a query amino acid's identity at the query amino acid position is systematically varied from one peptide pool to the next peptide pool within

15   the test set of peptide pools;

            (c) each degenerate amino acid position within every peptide of every peptide pool is occupied by an amino acid selected from a defined mixture of amino acids; and

            (d) the query amino acid position is not adjacent to the

20   phosphorylatable amino acid position in any peptide pool of the test set.

2.   The test set of claim 1, wherein at least one degenerate position in each peptide pool in the test set is occupied by a defined mixture of more than

25   five amino acids.

3.   The test set of claim 1, wherein the defined mixture comprises all natural amino acids except cysteine.

4.   The test set of claim 1, wherein each amino acid's relative abundance in the defined mixture is approximately that amino acid's relative

30   abundance in the human proteome.

5.   The test set of claim 1, wherein the defined mixture of amino acids comprises arginine.

6.   The test set of claim 1, wherein the test set has at least four peptide pools and each of the four peptide pools have a different query amino acid.

7.   The test set of claim 1, wherein the query amino acid position is two positions N-terminal to the phosphorylatable amino acid position.

8.   The test set of claim 1, wherein the query amino acid position is two positions C-terminal to the phosphorylatable amino acid position.

9.   The test set of claim 1, wherein one query amino acid is arginine.

10.   The test set of claim 1, wherein each peptide pool is a soluble mixture of peptides.

11.   The test set of claim 1, wherein substantially every peptide in each peptide pool is linked to biotin.

12.   The test set of claim 1, wherein substantially every peptide in each peptide pool is attached to a solid support

13.   The test set of claim 1 which also comprises at least one anchor amino acid position, and wherein:

        (a)   each anchor amino acid position is at a defined position relative to the phosphorylatable amino acid position within every peptide of every peptide pool and each anchor amino acid position has an identical anchor amino acid at that anchor amino acid position within every peptide of every peptide pool; and

        (b)   the query amino acid position is not adjacent to an anchor amino acid position in any peptide pool of the test set.

14.   The test set of claim 13, wherein at least one anchor amino acid is arginine.

15.   The test set of claim 13, wherein an anchor amino acid position is located one position C-terminal to the phosphorylatable amino acid position.

16.   The test set of claim 13, wherein an anchor amino acid position is located three positions N-terminal to the phosphorylatable amino acid position.

17.   The test set of claim 16, wherein arginine is the anchor amino acid at the anchor amino acid position located three positions N-terminal to the phosphorylatable amino acid position.

18.   The test set of claim 13, wherein every peptide in each of the peptide pools comprises less than four anchor amino acids

19.   A test set for characterizing substrate specificities of kinases comprising at least two peptide pools, wherein every peptide in each of the peptide

pools comprises one phosphorylatable amino acid position, one query amino acid, and at least one degenerate amino acid position, and wherein:

(a) each peptide of every peptide pool has an identical phosphorylatable amino acid that can be phosphorylated by a kinase at the phosphorylatable amino acid position;

(b) every peptide of every peptide pool has an identical query amino acid but the position of the query amino acid relative to the phosphorylatable amino acid position is systematically varied from one peptide pool to the next peptide pool within the test set of peptide pools; and

(c) each degenerate amino acid position within every peptide of every peptide pool is occupied by an amino acid from a defined mixture of amino acids.

20. The test set of claim 19, wherein the query amino acid is arginine.

21. The test set of claim 19, wherein each peptide of every peptide pool has at least one anchor amino acid position that is at a defined position relative to the phosphorylatable amino acid position, and wherein each anchor amino acid position of peptides within a peptide pool has an identical anchor amino acid at that anchor amino acid position.

22. The test set of claim 21, wherein the anchor amino acid is arginine and the anchor amino acid position is two positions N-terminal to the phosphorylatable amino acid position.

23. A test set of peptides for characterizing kinase substrate specificity comprising at least 50 separate peptides, each peptide consisting essentially of a sequence of between 6 and 30 amino acids, wherein each peptide sequence is different from every other peptide sequence, and wherein at least 50 peptides comprise two or more arginines within 6 amino acid positions of a serine or threonine.

24. The test set of claim of 23, wherein the test set has at least 96 separate peptides that comprise two or more arginines within 6 amino acid positions of a serine or threonine.

25. The test set of claim of 23, wherein at least half of the peptides comprise two or more arginines within 6 residues of a serine or threonine.

26.   The test set of claim of 23, wherein at least 50 peptides comprise two or more arginines but two of said arginines are not located 3 positions N-terminal to the serine or threonine.

27.   The test set of claim of 23, wherein at least 50 peptides comprise three or more arginine residues within 6 residues of a serine or threonine.

28.   The test set of claim of 23, wherein the at least 50 peptides further comprise one or more lysine residues within 6 residues of a serine or threonine.

29.   The test set of claim of 23, wherein substantially every peptide in the set corresponds to a peptidyl sequence in a mammalian protein and the peptidyl sequence is within 30 amino acids of the protein's N-terminus or C-terminus

30.   A peptide set consisting essentially of two or more pools of peptides, wherein each pool comprises peptides having substantially identical peptide sequences and the peptide sequences in each pool are selected from the group consisting essentially of SEQ ID NO: 76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-516 or 517.

31.   An isolated peptide consisting essentially of SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-516 or 517.

32.   The peptide of claim 31, wherein a serine or threonine in the peptide is phosphorylated.

33.   A binding entity whose binding differentiates between a peptide having any one of SEQ ID NO:76, 81, 82, 87, 89-92, 94, 97-99, 102, 104, 105, 108, 110, 112, 113, 121, 124, 127-129, 131-134, 136, 139, 143, 144, 149, 151-154, 160, 163-171, 173-177, 179, 182-192, 196-206, 208-211, 213-216, 474-517, and the peptide after phosphorylation by protein kinase C theta; wherein the binding entity has substantially no binding to a phosphorylated peptide having SEQ ID NO: 229 (WKN-pS-IRH).

34. The binding entity of claim 33, wherein the binding entity binds with greater affinity to the peptide after phosphorylation than before phosphorylation.

35. The binding entity of claim 33, wherein the binding entity binds with greater affinity to the peptide before phosphorylation than after phosphorylation.

36. The binding entity of claim 33, wherein the binding entity is an antibody, an antibody fragment or a mixture thereof.

37. The binding entity of claim 33, wherein the peptide is part of a mammalian protein.

38. The binding entity of claim 37, wherein the peptide's sequence is within 30 amino acids of the protein's N-terminus or C-terminus of said protein.

39. The binding entity of claim 38, wherein the peptide comprises any one of SEQ ID NO: 89, 102, 110, 112, 127, 177, 182, 209, 474-488 or 489.

40. The binding entity of claim 38 where the peptide comprises any one of SEQ ID NO: 173, 185, 192, 196, 200, 490-491 or 492.

41. The binding entity of claim 33 whose binding further differentiates between a phosphorylated peptide having any one of SEQ ID NO: 298, 301-324, 326-347, 349-400, 402-410, 412-473, 571-643 or 644, and a non-phosphorylated peptide that differs from the phosphorylated peptide by substitution of Ser for the pSer or substitution of a Thr for the pThr.

42. The binding entity of claim 41, wherein the phosphorylated peptide comprises any one of SEQ ID: 298, 320, 324, 350, 351, 366, 388, 394, 398, 402, 418, 464, 571-595 or 596.

43. The binding entity of claim 41, wherein the phosphorylated peptide comprises any one of SEQ ID: 301, 310, 317, 322, 344, 352, 371, 406, 597-599 or 600.

44. The binding entity of claim 41, wherein the phosphorylated peptide comprises SEQ ID NO:298.

45. The binding entity of claim 41, wherein the phosphorylated peptide comprises SEQ ID NO:313 or 314.

46. The binding entity of claim 41, wherein the phosphorylated peptide comprises SEQ ID NO:361 or 362.

47. A method for characterizing substrate specificities of kinases comprising:

(a) contacting each peptide pool in at least two test sets of

peptide pools with ATP and a kinase;

(b) quantifying the amount of phosphorylation in each peptide

pool; and

5   (c) comparing the amount of phosphorylation in each peptide

pool with the amount of phosphorylation in at least one

other peptide pool;

wherein substantially every peptide in each of the peptide pools

comprises one phosphorylatable amino acid position, one query

10   amino acid position, and at least one degenerate amino acid

position; and wherein

each peptide of every peptide pool has an identical

phosphorylatable amino acid that can be phosphorylated by a

kinase at the phosphorylatable amino acid position;

15   the query amino acid position is at a defined position relative to the

phosphorylatable amino acid position within every peptide of

every peptide pool but a query amino acid's identity at the query

amino acid position is systematically varied from one peptide

pool to the next peptide pool within the test set of peptide pools;

20   and

each degenerate amino acid position within every peptide of every

peptide pool is occupied by an amino acid from a defined mixture

of amino acids.

48.   The method of claim 47, wherein quantifying the amount of

25   phosphorylation comprises determining a total amount of labeled

phosphate incorporated into each peptide pool.

49.   The method of claim 47, wherein quantifying the amount of

phosphorylation comprises determining a total amount of phosphorylated

peptide in each peptide pool with an antibody specific for a

30   phosphorylated peptide.

50.   A method for visual display of amino acid or nucleotide sequence

preferences comprising a series of stacks of single letter symbols for

amino acids or nucleotides, wherein

(a) each stack represents a position in a peptide or a nucleic acid sequence;

(b) each symbol's height is proportional to the absolute value of a quantitative parameter that is positive for favored amino acids or nucleotides and negative for disfavored amino acids or nucleotides; and

(c) each symbol's position within the stack is sorted from bottom to top in ascending value by the quantitative parameter.

Fig 1. Examples of two test sets of peptides and results with PKC-theta

A. P+1 test set

	Peptide Sequence	Symbolic Representation: ddddRdd-S-?dd	CPM	Ratio to Mean	Log Score
1	ddddRdd-S–	Ddd	67* ± 3	0.4	-1.3 ± 0.03
2	ddddRdd-S–	Ndd	88 ± 3	0.5	-0.9 ± 0.04
3	ddddRdd-S–	Qdd	158 ± 4	1.0	0.0 ± 0.04
4	ddddRdd-S–	Rdd	474 ± 9	2.9	1.6 ± 0.04
5	ddddRdd-S–	Kdd	406 ± 18	2.5	1.3 ± 0.08
6	ddddRdd-S–	Hdd	163 ± 10	1.0	0.0 ± 0.09
7	ddddRdd-S–	Sdd	136 ± 3	0.8	-0.2 ± 0.03
8	ddddRdd-S–	Pdd	90 ± 3	0.6	-0.8 ± 0.07
9	ddddRdd-S–	Ldd	243 ± 18	1.5	0.6 ± 0.11
10	ddddRdd-S–	Fdd	461 ± 8	2.9	1.5 ± 0.03
11	ddddRdd-S–	Wdd	135 ± 6	0.8	-0.3 ± 0.07
12	ddddRdd-S–	Gdd	93 ± 2	0.6	-0.8 ± 0.03
13	ddddRdd-S–	Add	97 ± 5	0.6	-0.7 ± 0.10

B. P+2 test set

	Peptide Sequence	Symbolic Representation: ddddRdd-S-F?d	CPM	Ratio to Mean	Log Score
1	ddddRdd-S-F	Dd	186 ± 11	0.5	-1.1 ± 0.08
2	ddddRdd-S-F	Nd	696 ± 22	1.7	0.8 ± 0.04
3	ddddRdd-S-F	Qd	229 ± 7	0.6	-0.8 ± 0.05
4	ddddRdd-S-F	Rd	1248 ± 38	3.1	1.6 ± 0.05
5	ddddRdd-S-F	Kd	1027 ± 25	2.5	1.3 ± 0.05
6	ddddRdd-S-F	Hd	696 ± 35	1.7	0.8 ± 0.06
7	ddddRdd-S-F	Sd	431 ± 10	1.1	0.1 ± 0.05
8	ddddRdd-S-F	Pd	148 ± 3	0.4	-1.5 ± 0.03
9	ddddRdd-S-F	Ld	489 ± 5	1.2	0.3 ± 0.01
10	ddddRdd-S-F	Fd	543 ± 19	1.3	0.4 ± 0.07
11	ddddRdd-S-F	Wd	459 ± 3	1.1	0.2 ± 0.01
12	ddddRdd-S-F	Gd	195 ± 2	0.5	-1.1 ± 0.01
13	ddddRdd-S-F	Ad	201 ± 7	0.5	-1.0 ± 0.07

*Values for substantially favored residues are bold underlined; values for substantially disfavored residues are bold without underline.

Fig 2: A superset of test sets designed for analysis of PKC specificity from P-4 to P+3

Fig 3: Raw Counts For In Vitro phosphorylation by PKC-theta of a collection of peptides designed for analysis of PKC specific

Residue	-4	-3	-2	-1	P0	+1	+2	+3
D	277*	78	419	294		67	186	263
N	422	153	477	651		88	696	610
Q	411	200	493	472		158	229	358
R	2072	461	1746	776		474	1248	843
K	966	117	918	846		406	1027	832
H	573	117	1330	461		163	696	506
S	433	90	389	371	461	136	431	304
P	439	93	458	376		90	148	226
L	640	85	347	413		243	489	715
F	741	246	243	348		461	543	647
W	721	151	301	291		135	459	470
G	517	127	449	672		93	195	409
A	380	83	362	678		97	201	711
T					402			
Geo Mean For Set	572	133	509	481	430	161	406	488

*Raw counts for substantially favored residues are bold underlined; raw counts for substantially disfavored residues are bold without underline.

Fig 4: Ratio to Mean for PKC-theta

Residue	-4	-3	-2	-1	P0	+1	+2	+3
D	0.5*	0.6	0.8	0.6		0.4	0.5	0.5
N	0.7	1.1	0.9	1.4		0.5	1.7	1.3
Q	0.7	1.5	1.0	1.0		1.0	0.6	0.7
R	3.6	3.5	3.4	1.6		2.9	3.1	1.7
K	1.7	0.9	1.8	1.8		2.5	2.5	1.7
H	1.0	0.9	2.6	1.0		1.0	1.7	1.0
S	0.8	0.7	0.8	0.8	1.1	0.8	1.1	0.6
P	0.8	0.7	0.9	0.8		0.6	0.4	0.5
L	1.1	0.6	0.7	0.9		1.5	1.2	1.5
F	1.3	1.9	0.5	0.7		2.9	1.3	1.3
W	1.3	1.1	0.6	0.6		0.8	1.1	1.0
G	0.9	1.0	0.9	1.4		0.6	0.5	0.8
A	0.7	0.6	0.7	1.4		0.6	0.5	1.5
T					0.9			

*Ratio to mean for substantially favored residues are bold underlined; ratio to mean for substantially disfavored residues are bold without underline.

Fig 5: Position-specific scoring matrix for PKC-theta

	Residue	-4	-3	-2	-1	P0	+1	+2	+3
Experimentally determined	D	-1.0*	-0.8	-0.3	-0.7		-1.3	-1.1	-0.9
	N	-0.4	0.2	-0.1	0.4		-0.9	0.8	0.3
	Q	-0.5	0.6	0.0	0.0		0.0	-0.8	-0.4
	R	1.9	1.8	1.8	0.7		1.6	1.6	0.8
	K	0.8	-0.2	0.9	0.8		1.3	1.3	0.8
	H	0.0	-0.2	1.4	-0.1		0.0	0.8	0.1
	S	-0.4	-0.6	-0.4	-0.4	0.1	-0.2	0.1	-0.7
	P	-0.4	-0.5	-0.2	-0.4		-0.8	-1.5	-1.1
	L	0.2	-0.7	-0.6	-0.2		0.6	0.3	0.6
	F	0.4	0.9	-1.1	-0.5		1.5	0.4	0.4
	W	0.3	0.2	-0.8	-0.7		-0.3	0.2	-0.1
	G	-0.1	-0.1	-0.2	0.5		-0.8	-1.1	-0.3
	A	-0.6	-0.7	-0.5	0.5		-0.7	-1.0	0.5
	T					-0.1			
Extrapolated	T	-0.4	-0.6	-0.4	-0.4		-0.2	0.1	-0.7
	E	-1.0	-0.8	-0.3	-0.7		-1.3	-1.1	-0.9
	I	0.2	-0.7	-0.6	-0.2		0.6	0.3	0.6
	M	0.2	-0.7	-0.6	-0.2		0.6	0.3	0.6
	V	0.2	-0.7	-0.6	-0.2		0.6	0.3	0.6
	Y	0.4	0.9	-1.1	-0.5		1.5	0.4	0.4
	C	-0.6	-0.7	-0.5	0.5		-0.7	-1.0	0.5

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

Fig 6. A superset of degenerate peptides designed to extend analysis of PKC specificity

Superset:		Superset A					
Anchored Residues		Rxx-S-F					
Set		Set P-7	Set P-6	Set P-5	Set P+4	Set P+5	Set P+6
Position		P-7	P-6	P-5	P+4	P+5	P+6
Symbolic Representation of Set		dddd?Rdd-S-Fdd	dddddR?d-S-Fdd	ddddddRd?-S-Fdd	ddRdd-S-F?dddd	ddRdd-S-F?dddd	ddRdd-S-Fd?dd
Res	**Other residues in class**						
D	E	9D999R99-S-F99	99D99R99-S-F99	99D9DR99-S-F99	99R99-S-F99D99	99R99-S-F999D9	99R99-S-F9999D
N		9N999R99-S-F99	99N99R99-S-F99	99N9DR99-S-F99	99R99-S-F99N99	99R99-S-F999N9	99R99-S-F9999N
Q		9Q999R99-S-F99	99Q99R99-S-F99	99Q9DR99-S-F99	99R99-S-F99Q99	99R99-S-F999Q9	99R99-S-F9999Q
R		9R999R99-S-F99	99R99R99-S-F99	99R9DR99-S-F99	99R99-S-F99R99	99R99-S-F999R9	99R99-S-F9999R
K		9K999R99-S-F99	99K99R99-S-F99	99K9DR99-S-F99	99R99-S-F99K99	99R99-S-F999K9	99R99-S-F9999K
H	E	9H999R99-S-F99	99E99R99-S-F99	99E9SR99-S-F99	99R99-S-F99H99	99R99-S-F999H9	99R99-S-F9999H
S		9S999R99-S-F99	99S99R99-S-F99	99S9SR99-S-F99	99R99-S-F99S99	99R99-S-F999S9	99R99-S-F9999S
P		9P999R99-S-F99	99P99R99-S-F99	99P9PR99-S-F99	99R99-S-F99P99	99R99-S-F999P9	99R99-S-F9999P
L	IMV	9L999R99-S-F99	99L99R99-S-F99	99L9LR99-S-F99	99R99-S-F99L99	99R99-S-F999L9	99R99-S-F9999L
F	Y	9F999R99-S-F99	99F99R99-S-F99	99F9FR99-S-F99	99R99-S-F99F99	99R99-S-F999F9	99R99-S-F9999F
W		9W999R99-S-F99	99W99R99-S-F99	99W9WR99-S-F99	99R99-S-F99W99	99R99-S-F999W9	99R99-S-F9999W
G		9G999R99-S-F99	99G99R99-S-F99	99G9GR99-S-F99	99R99-S-F99G99	99R99-S-F999G9	99R99-S-F9999G
A		9A999R99-S-F99	99A99R99-S-F99	99A9AR99-S-F99	99R99-S-F99A99	99R99-S-F999A9	99R99-S-F9999A
E							
H							
M							
Y							

Residues that vary within a variable position

Fig 7. Extension of Position-specific scoring matrix for PKC-theta

	Residue	-7	-6	-5	+4	+5	+6
Experimentally determined	D	-0.9*	-0.5	-0.9	-0.3	-0.4	-0.4
	E						
	N	0.1	-0.1	-0.2	-0.1	**0.6**	0.1
	Q	-0.2	0.0	-0.2	-0.2	0.0	-0.3
	R	**0.8**	**1.1**	**0.8**	**0.9**	0.0	**0.7**
	K	0.4	**0.5**	**0.5**	**0.9**	0.5	**0.7**
	H	-0.1	-0.1	0.1	**0.5**	0.2	0.0
	S	0.0	0.1	0.4	0.4	-0.2	0.1
	T						
	P	0.5	0.3	-0.2	0.0	0.0	0.1
	L	0.0	-0.5	0.5	-0.6	0.3	-0.3
	M						
	F	0.0	-0.3	0.1	-0.3	-0.1	0.1
	Y						
	W	-0.9	-0.4	-0.5	-0.8	-0.5	-1.1
	G	0.4	0.1	-0.2	0.0	0.0	0.1
	A	0.0	-0.2	-0.2	-0.3	-0.4	0.1

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

Fig 8. Sequence Logo vs PSSM Logo
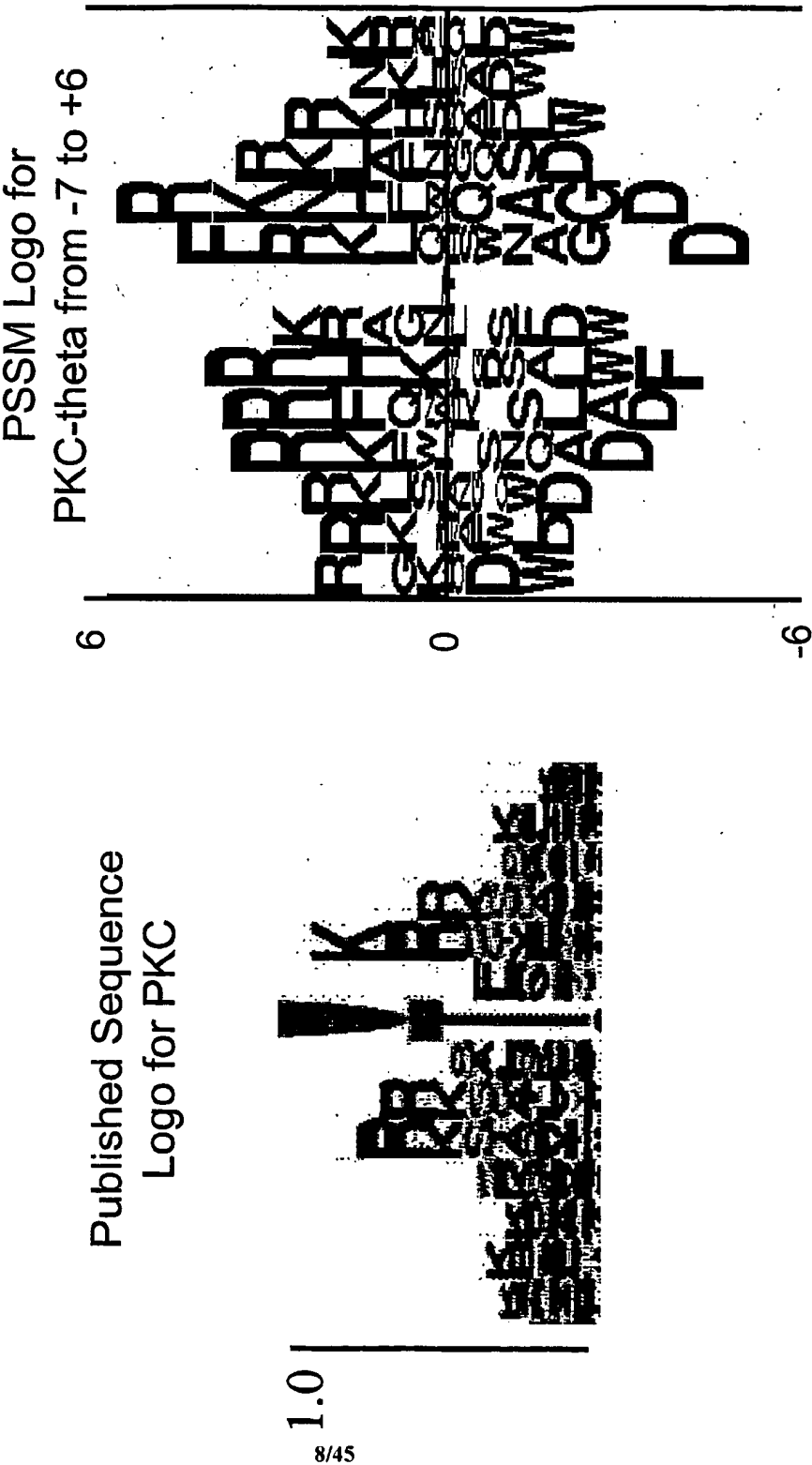
Fig 9. Testing our predictions for PKC-theta and Scansite prediction for PKC-delta against results for PKC-delta
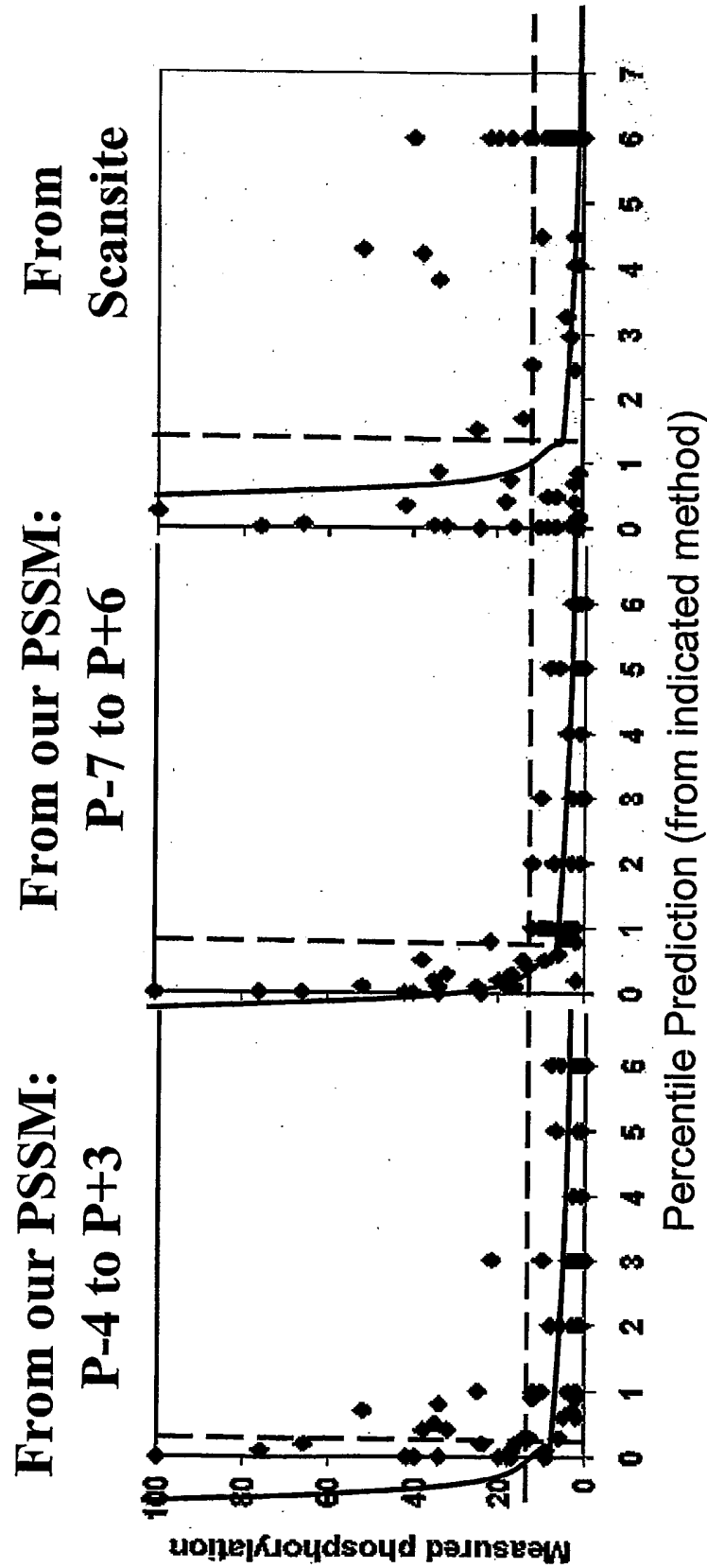
Fig 10: Comparison of our scoring with Scansite scoring for PKC-delta

Criteria:

Positive score: better than 1st percentile

Positive phosphorylation: better than 10% of max

Ours

		Measured activity	
		Positive	Negative
Prediction	Positive	21	4
score	Negative	2	48
Total Scored		75	
Sensitivity		92	
Specificity		91	

Scansite

		Measured activity	
		Positive	Negative
Prediction	Positive	10	16
score	Negative	12	37
Total Scored		75	
Sensitivity		70	
Specificity		45	

Fig 11. Prospective validation of scoring for PKC-theta PSSM
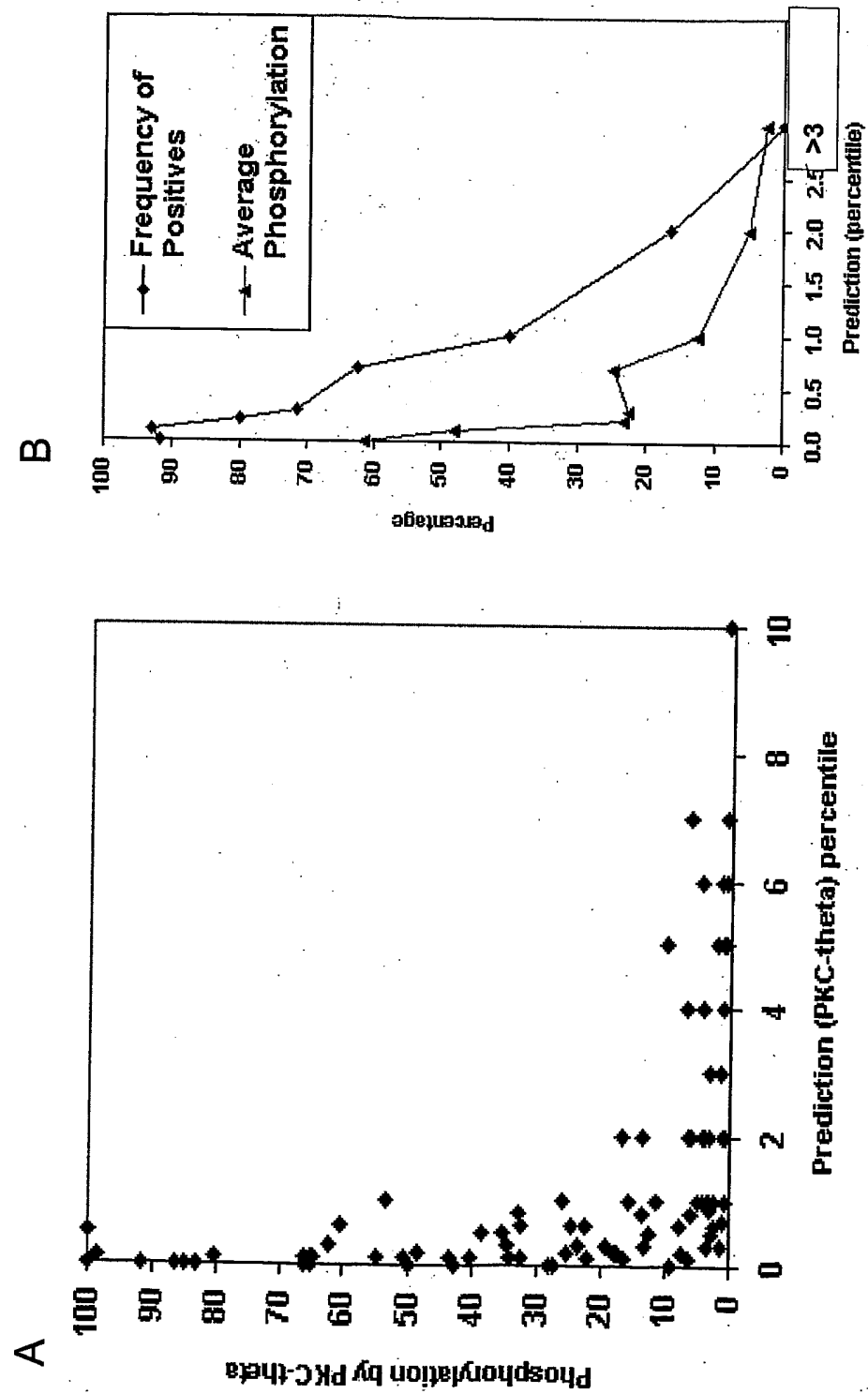
Fig 12. The d??R??S????d superset of test sets with 1 anchor position

Fig 13. PSSM Logo from analysis of d??R??S????d superset of peptides with AKT1
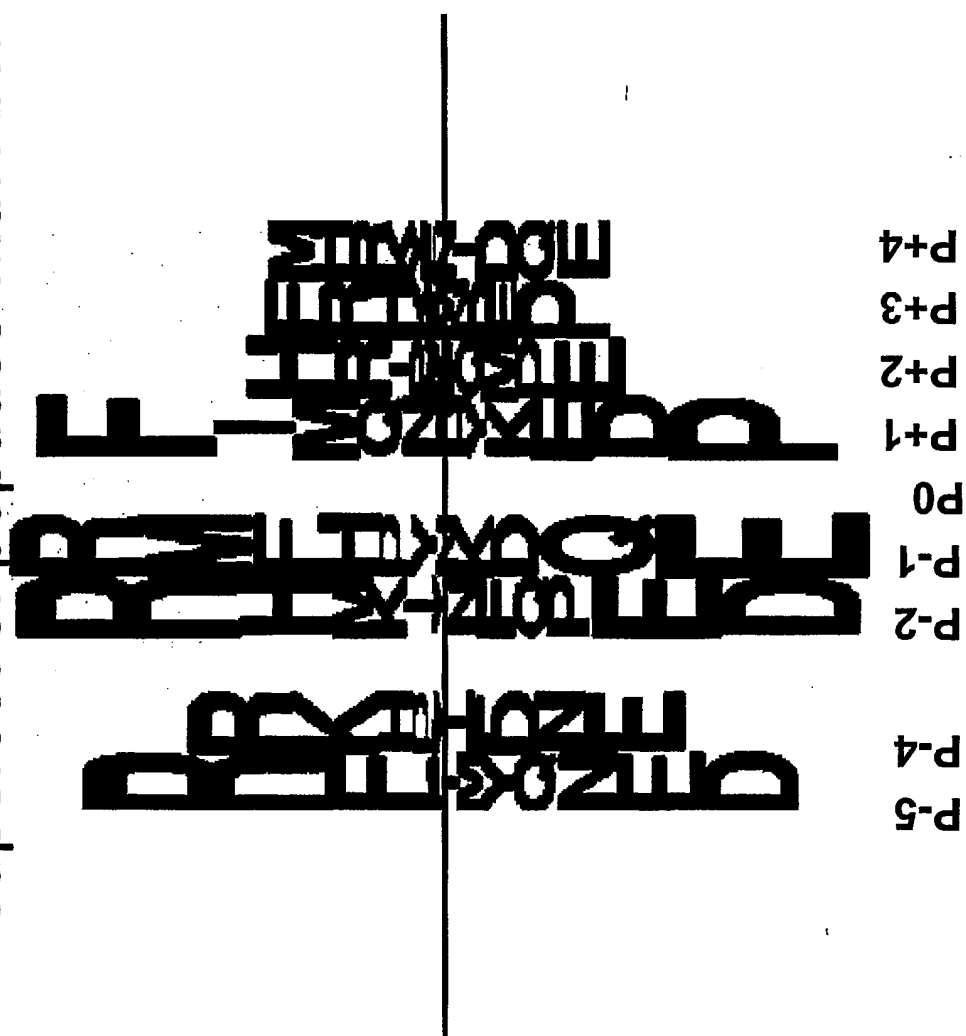
Fig 14: Issues related to abundance of residues in degenerate positions

Amino Acid		Hydrophobicity Score	Target percentages			Estimated incorporation efficiency	Reagent parts per hundred	Percent by sequencing
			Plan 1	Plan 2	Plan 3			
Code	Name		Equal Abundance	Abundance In Human Proteome	Abundance in PKC substrates			
A	Alanine	0.31	5	7.1	6.6	1.1	8	15.8
C	Cysteine	1.54	5	2.4	0.8		0	
D	Aspartic acid	-0.77	5	4.5	5.8	1.0	6	5.9
E	Glutamic	-0.64	5	6.8	7.3	1.2	6	6.1
F	Phenylalanin	1.79	5	3.6	3.2	1.2	4	3.8
G	Glycine	0	5	6.9	7	1.0	8	9.2
H	Histidine	0.13	5	2.7	1.7	0.9	4	2.7
I	Isoleucine	1.8	5	4.2	3.3	0.7	4	1.6
K	Lysine	-0.99	5	5.7	7.1	1.1	6	8.9
L	Leucine	1.7	5	9.8	6.8	0.9	6	9.1
M	Methionine	1.23	5	2.2	1.9	1.1	2	2.7
N	Asparagine	-0.6	5	3.5	2.9	1.0	4	3.0
P	Proline	0.72	5	6.5	7.6	1.0	6	3.7
Q	Glutamine	-0.22	5	4.8	2.8	1.2	6	3.9
R	Arginine	-1.01	5	5.8	8.9	1.1	6	5.9
S	Serine	-0.04	5	8.4	12.2	1.1	8	4.6
T	Threonine	0.26	5	5.4	5.9	0.8	6	3.1
V	Valine	1.22	5	5.9	4.3	0.8	4	3.5
W	Tryptophane	2.25	5	1.3	0.3		2	0.2
Y	Tyrosine	0.96	5	2.5	3.6	0.9	4	6.2

FIG. 15: Detection of specific phosphorylation of SHP-1 by Western blot with pPKC antibody which is augmented following stimulation by the T-cell receptor
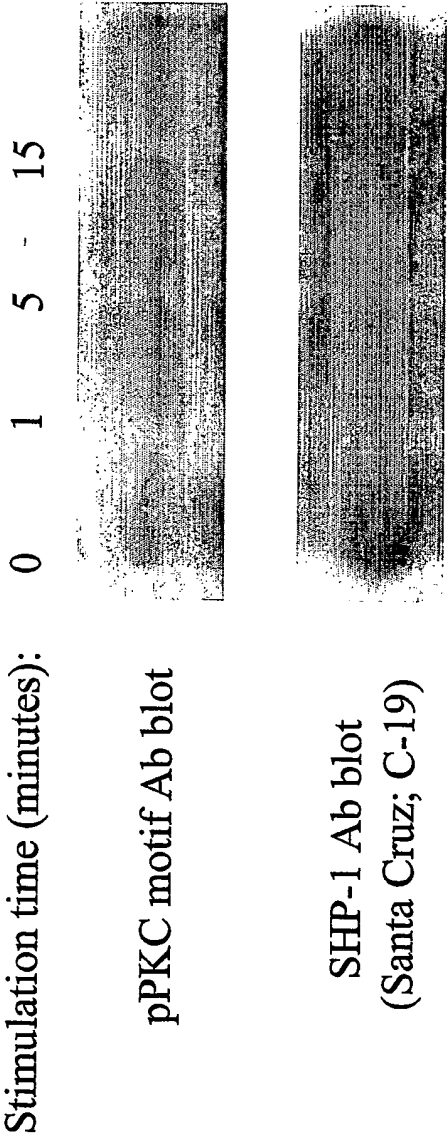
Scores derived from different test sets tested at different times are reproducible and scores extrapolated for untested residues can be adequate

FIG. 17

FIG. 16

Residue	Test Set	
	#1	#2
D	-1.3*	-1.3
E		-0.9
N	-0.9	-0.5
Q	0.0	0.5
R	1.6	1.2
K	1.3	0.7
H	0.0	-0.2
S	-0.2	-0.3
T		0.2
P	-0.8	-1.0
L	0.6	0.7
M		0.8
F	1.5	1.4
Y		0.5
W	-0.3	-0.4
G	-0.8	-0.7
A	-0.7	-0.6

Experimentally determined

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.
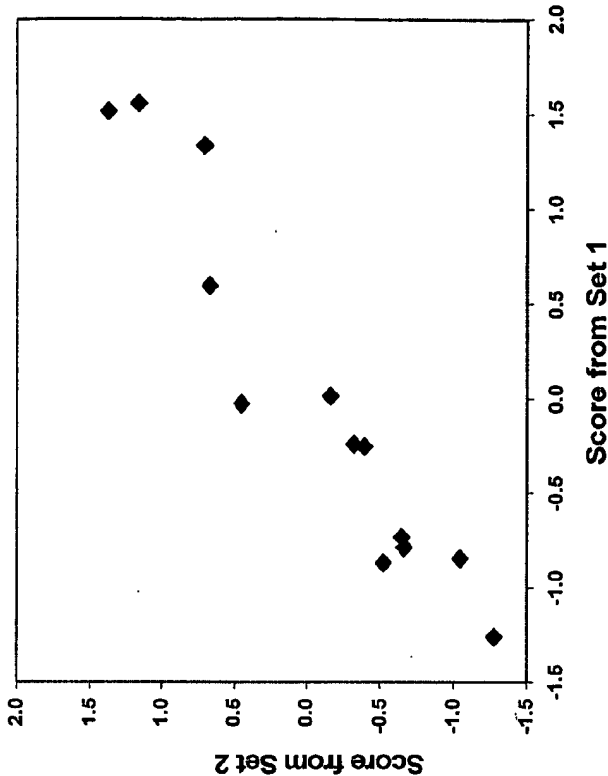
Fig 18: Scoring a peptide

Protein	Sequence Scored	Residue Number	-7	-6	-5	-4	-3	-2	-1	P0	+1	+2	+3	+4	+5	+6	Total Raw Score	Prediction Rank
MARCKS	KKKKKRFS-FKKSFK	159	K	K	K	K	K	R	F	S	F	K	K	S	F	K	7.9	0.02
			0.6	0.7	0.6	0.8	1.1	1.2	-0.1	0.0	0.9	1.2	0.6	0.3	-0.1	0.5		

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

Fig 19. Distribution of PKC scores sites in proteins encoded by 15651 human genes
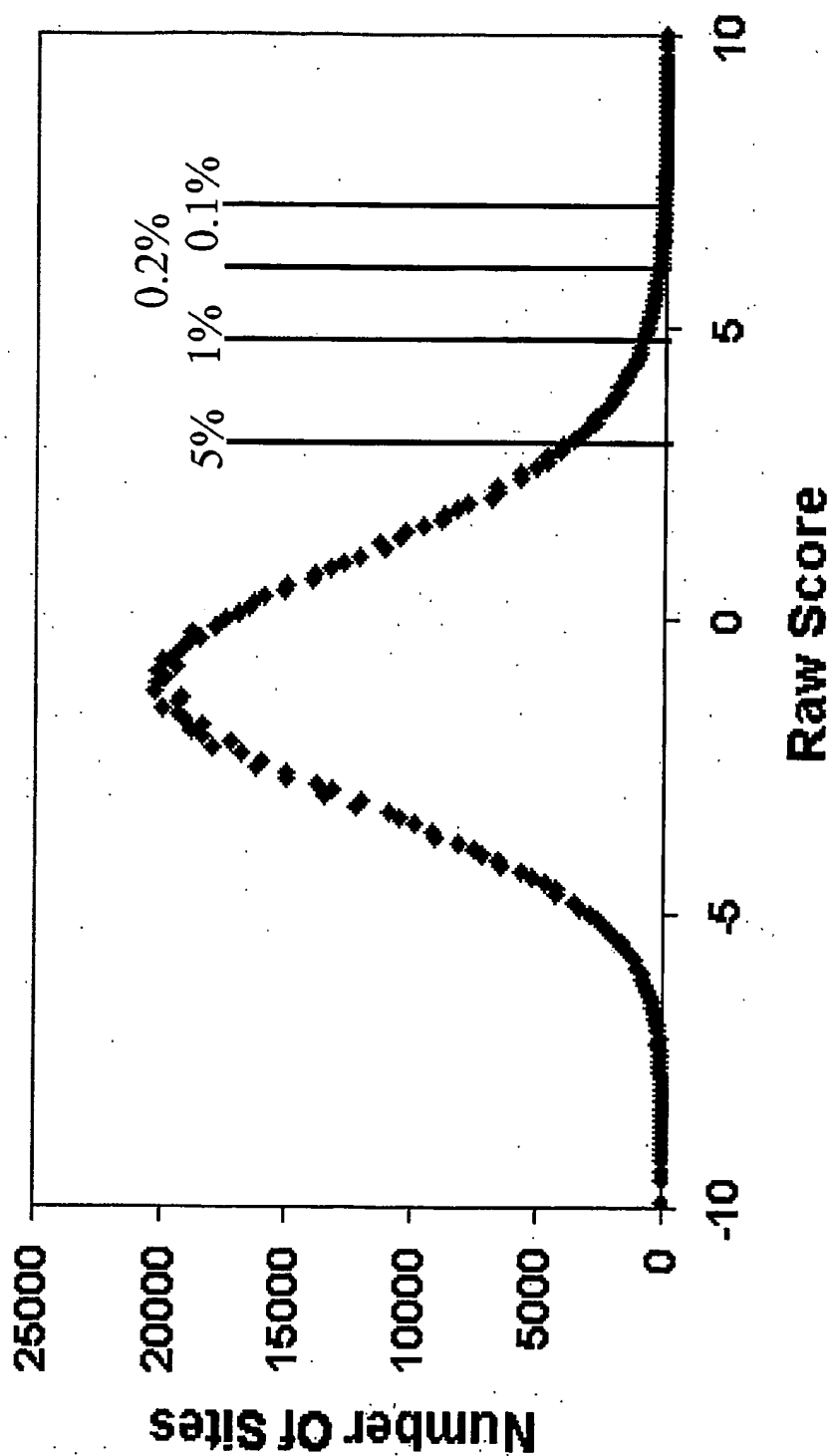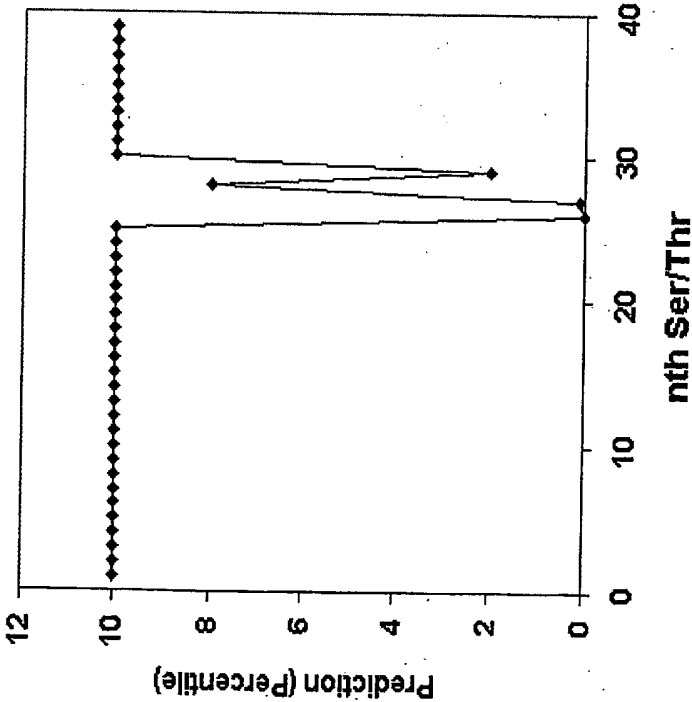
Fig 20: The PKC site prediction algorithm correctly predicts sites on MARCKS

Order	Sequence	Residue Number	Percentile Predicted		
			PKC Theta	AKT1	PKA
23	KAEDGATP-S-PSNETPKK	145	90	40	100
24	EDGATPSP-S-NETPKKKK	147	90	70	80
25	ATPSPSNE-T-PKKKKKKRF	150	10	80	100
25	PKKKKKRF-S-FKKSFKLS	159	**0**	**0.9**	10
27	KKRFSFKK-S-FKLSGFSF	163	**0.1**	10	10
28	SFKKSFKL-S-GFSFKKNK	167	8	70	10
29	KSFKLSGF-S-FKKNKKEA	170	2	20	90
30	AAAEAGAA-S-GEQAAAPG	211	100	40	80
31	VAPEKPPA-S-DETKAAEE	252	70	40	80

nth Ser/Thr — Prediction (Percentile)

FIG. 21. High similarity in specificity between novel and classical PKC isoforms, but atypical PKC differs more and great divergence seen with AKT1 and PKA

Correlation (arithmetic)	alpha	delta	epsilon	zeta	AKT1	PKA
alpha	1.00	0.86	0.89	0.69	0.38	0.29
delta	0.86	1.00	0.91	0.73	0.63	0.35
epsilon	0.89	0.91	1.00	0.76	0.50	0.38
zeta	0.69	0.73	0.76	1.00	0.35	0.28
AKT1	0.38	0.63	0.50	0.35	1.00	0.51
PKA	0.29	0.35	0.38	0.28	0.51	1.00

Fig 22. Differences between PSSM Logos of different kinases analyzed with the same peptide collections

PSSM Logo for PKA from -7 to +6

PSSM Logo for PKC-zeta from -7 to +6
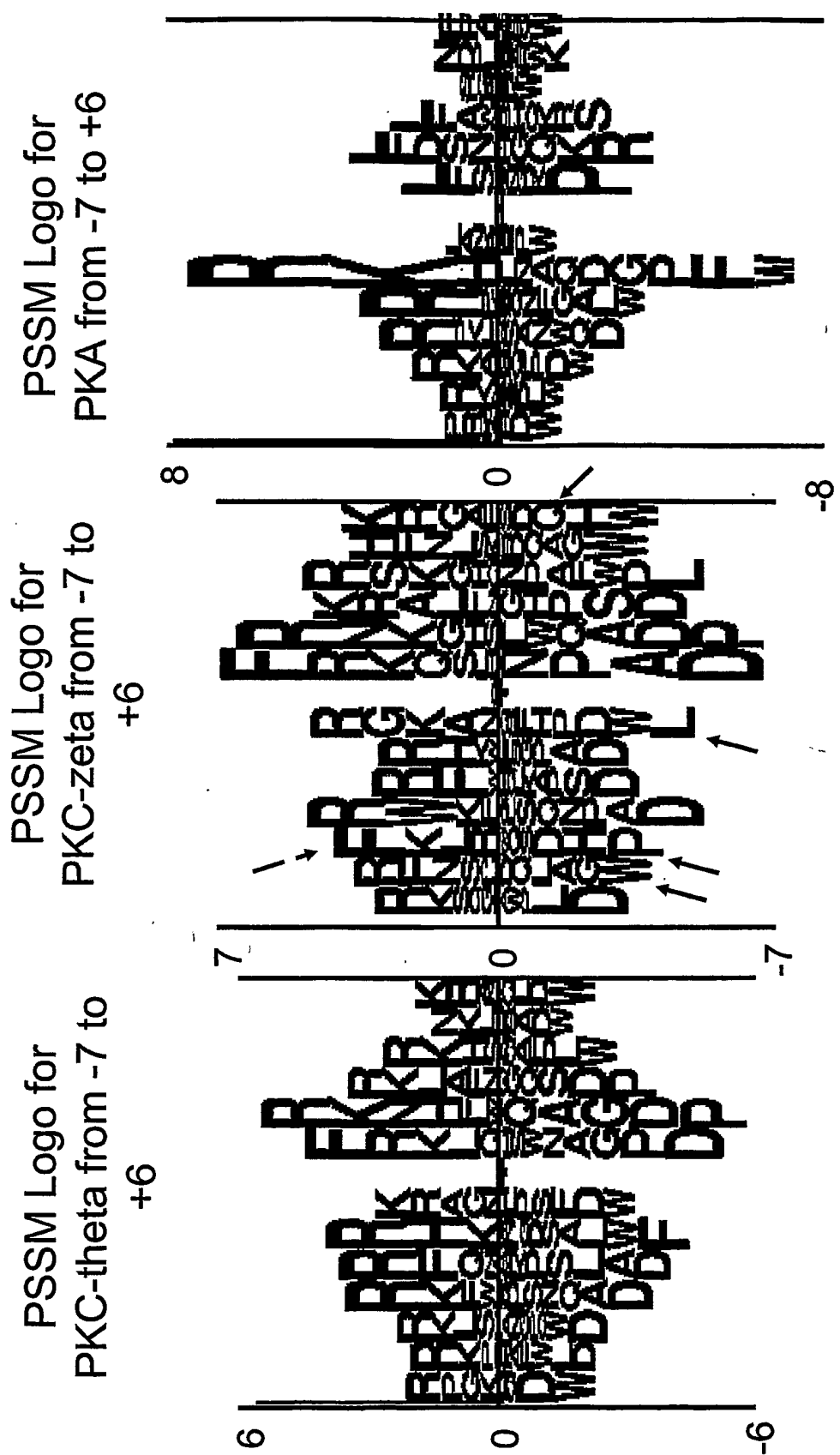
PSSM Logo for PKC-theta from -7 to +6

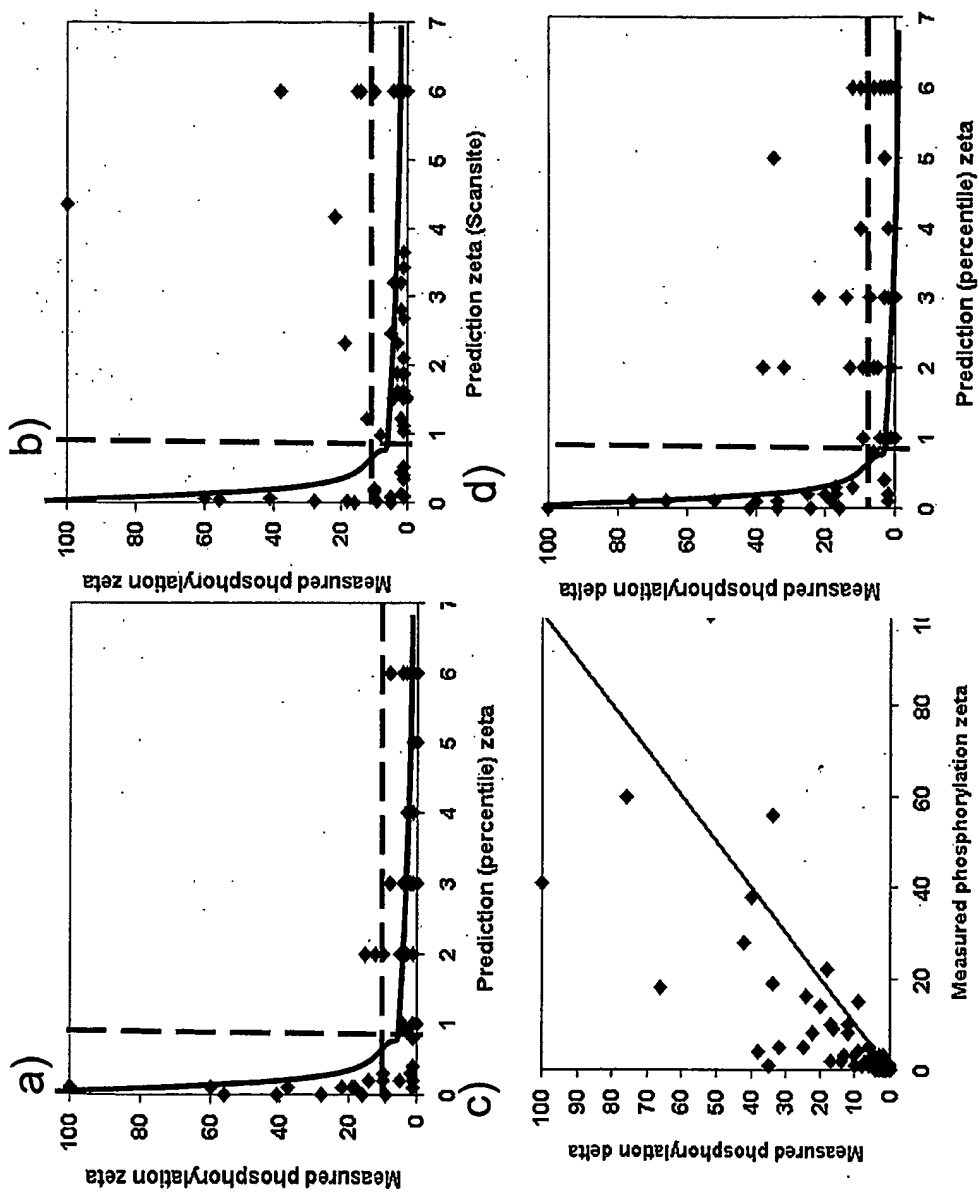Fig 23. Prediction for PKC-zeta predicts PKC-zeta better than PKC-delta

FIG. 24. Scoring changes in peptides that are less phosphorylated by PKC-zeta than PKC-delta

Peptide		Measured phosphorylation	Predicted percentile	7	6	5	4	3	2	1	0	+1	+2	+3	+4	+5	+6
1	residue			A	A	R	K	K	R	I	S	V	K	K	K	Q	E
	difference	34	-4.8	0.1	0.4	0.2	0	0	-0.1	1.4	-0.1	0.3	-0.6	-0.7	-0.1	0.5	0.1
	zeta	1	5.0	-0.1	-0.6	0.7	0.8	-0.2	1.9	-1.5	0.2	0.4	1.9	1.5	1.0	-0.5	-0.5
	delta	35	0.2	0.0	-0.2	0.9	0.8	-0.2	1.8	-0.2	0.1	0.7	1.3	0.8	0.9	0.0	-0.4
2	residue			K	I	P	K	R	P	G	S	V	H	R	T	P	S
	difference	34	-1.5	-0.2	0.1	1.2	0	-0.7	0.4	-1	-0.1	0.3	-0.3	-0.5	-0.6	0.2	0.6
	zeta	4	2.0	0.6	-0.6	-1.4	0.8	2.5	-0.6	1.5	0.2	0.4	1.1	1.3	1.0	-0.2	-0.4
	delta	38	0.5	0.4	-0.5	-0.2	0.8	1.8	-0.2	0.5	0.1	0.7	0.8	0.8	0.4	0.0	0.2
3	residue			G	A	P	P	R	R	S	S	I	R	N	A	H	
	difference	10	-1.6	0.1	0.4	1.2	0.3	-0.7	-0.1	-0.3	-0.1	0.3	-2	0	0.2	-0.8	
	zeta	3	2.0	0.3	-0.6	-1.4	-0.7	2.5	1.9	-0.1	0.2	0.4	3.6	0.3	-0.5	1.0	
	delta	13	0.4	0.4	-0.2	-0.2	-0.4	1.8	1.8	-0.4	0.1	0.7	1.6	0.3	-0.3	0.2	
4	residue			A	G	S	F	K	R	N	S	I	K	K	I	V	R
	difference	12	-2.5	0.2	0.7	0.9	0.0	0.1	0.2	-0.1	-0.1	0.4	0.1	0.3	0.7	-0.3	-0.3
	zeta	2	3.0	-0.2	-0.6	-0.5	0.4	-0.3	1.6	0.5	0.2	0.3	1.2	1.1	-1.3	0.6	
	delta	14	0.5	0.0	0.1	0.4	0.4	-0.2	1.8	0.4	0.1	0.7	1.3	0.8	-0.6	0.3	
5	residue				W	K	G	K	R	R	S	K	A	R	K	K	R
	difference	14	-2.2		0.9	-0.4	-0.1	0	-0.1	-0.9	-0.1	-0.3	0.6	-0.5	-0.1	0.2	-0.3
	zeta	8	3.0		-1.3	0.9	-0.1	-0.2	1.9	1.6	0.2	1.0	-1.6	1.3	1.0	0.3	1.0
	delta	22	0.8		-0.4	0.5	-0.2	-0.2	1.8	0.7	0.1	0.7	-1.0	0.8	0.9	0.5	0.7
6	residue			A	Q	I	V	K	R	A	S	L	K	R	G	K	Q
	difference	27	-1.7	0.1	0.5	-0.3	0.3	0	-0.1	-0.7	-0.1	0.3	-0.6	-0.5	-0.6	0.2	0.6
	zeta	5	2.0	-0.1	-0.5	0.8	-0.1	-0.2	1.9	1.2	0.2	0.4	1.9	1.3	0.5	0.3	-0.9
	delta	32	0.3	0.0	0.0	0.5	0.2	-0.2	1.8	0.5	0.1	0.7	1.3	0.8	0.0	0.5	-0.3

FIG. 25 Position-specific residue preference for PKA and PKG determined using the PKC superset

PKA

	-4	-3	-2	-1	P0	+1	+2	+3
D	0.6*	1.0	0.6	0.9		1.0	1.6	0.9
N	0.7	0.8	0.7	1.4		1.0	1.5	1.2
Q	0.7	0.8	0.6	1.0		0.9	0.6	0.8
R	3.8	8.7	12.1	1.1		1.2	0.5	0.7
K	1.3	1.1	8.9	1.1		1.0	0.5	0.8
H	1.1	0.9	2.1	0.9		1.1	0.8	0.6
S	0.8	1.1	0.9	0.9	0.9	1.0	1.5	0.9
P	1.3	1.2	0.5	0.8		0.3	1.1	1.9
L	1.2	0.6	0.7	1.3		1.9	1.7	1.7
F	1.1	0.7	0.3	1.0		1.8	1.7	1.2
W	0.7	0.6	0.3	0.6		1.2	1.0	1.0
G	0.9	0.7	0.9	1.0		0.9	0.7	1.5
A	0.8	0.7	0.6	1.2		0.9	1.0	
T					1.1			

PKG

	-4	-3	-2	-1	P0	+1	+2	+3
D	0.4	1.1	0.6	0.4		0.7	2.1	0.9
N	0.4	1.3	0.5	1.8		0.8	1.4	1.2
Q	0.6	1.0	0.9	1.2		0.7	0.3	0.8
R	5.1	3.2	28.2	1.9		2.4	0.4	0.9
K	2.2	3.6	11.6	2.0		1.6	0.4	0.8
H	1.3	1.0	0.6	1.5		0.7	1.2	0.2
S	0.6	0.9	1.0	0.8		0.8	1.5	0.7
P	1.0	0.9	0.6	0.5	0.9	0.5	0.5	0.9
L	1.1	0.6	0.5	1.2		2.6	2.1	3.0
F	1.1	0.9	0.4	2.2		1.5	2.8	2.5
W	0.6	0.5	0.5	0.7		1.3	1.2	1.5
G	1.5	0.4	0.6	0.3		0.4	0.7	1.1
A	1.0	0.6	0.4	0.8		1.1	1.5	1.6
T					1.1			

*Values expressed as ratio to mean. Values for favored residues are bold underlined; values for substantially disfavored residues are bold without underline.

Fig 26. Use of peptides to analyze mutant constructs and altered conditions for kinase assay
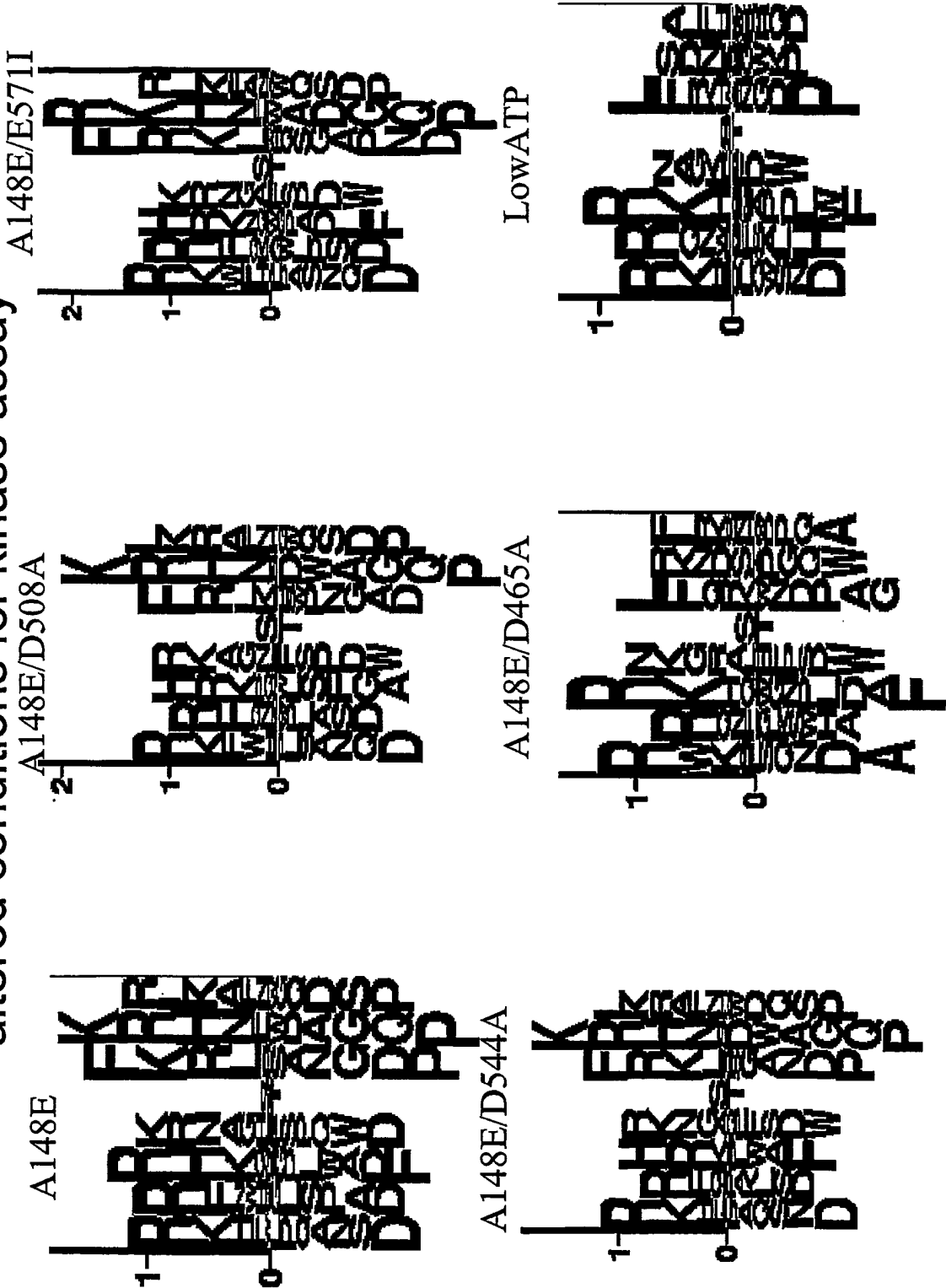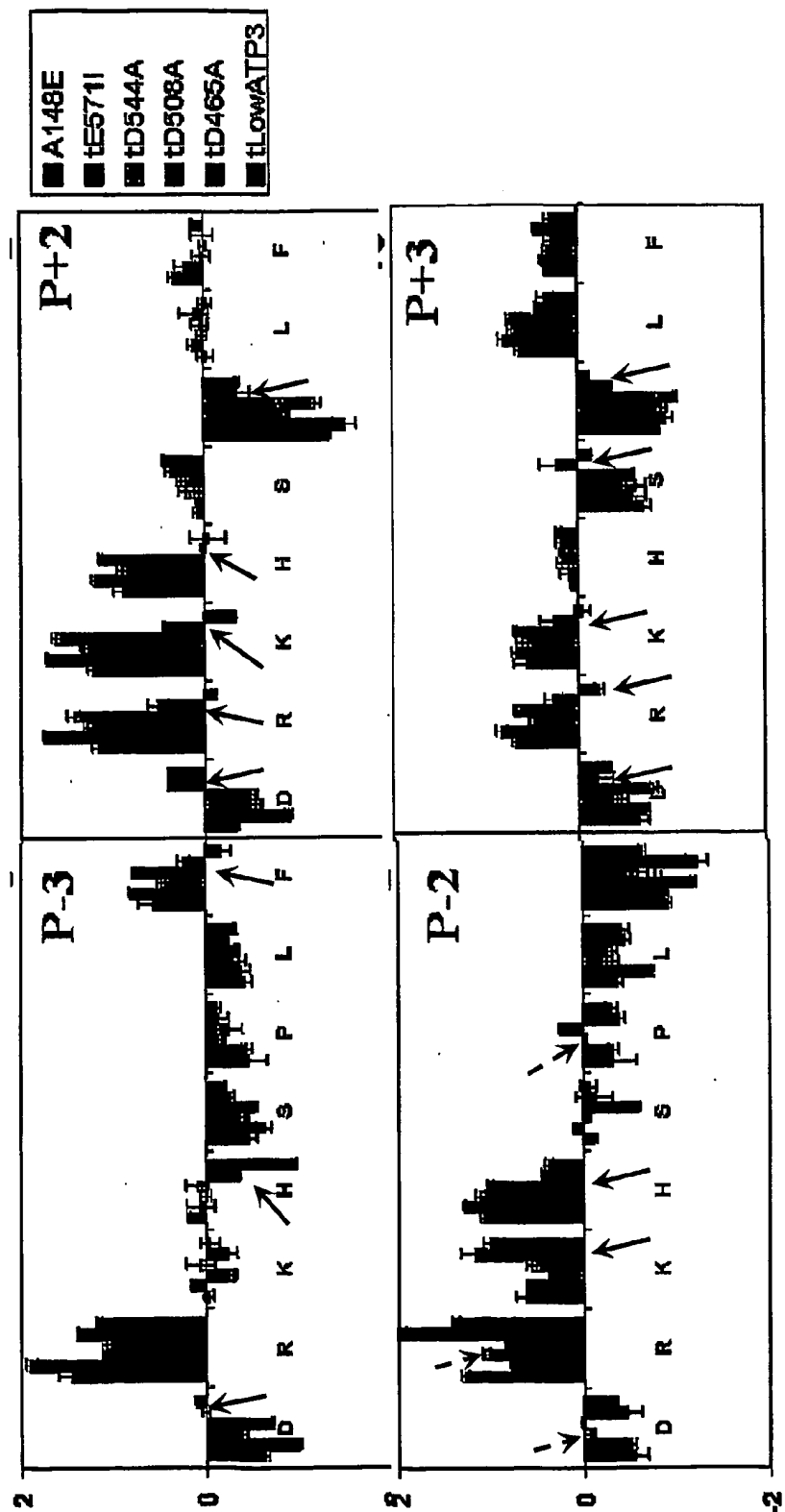
Fig 27. Details of changes observed with mutant
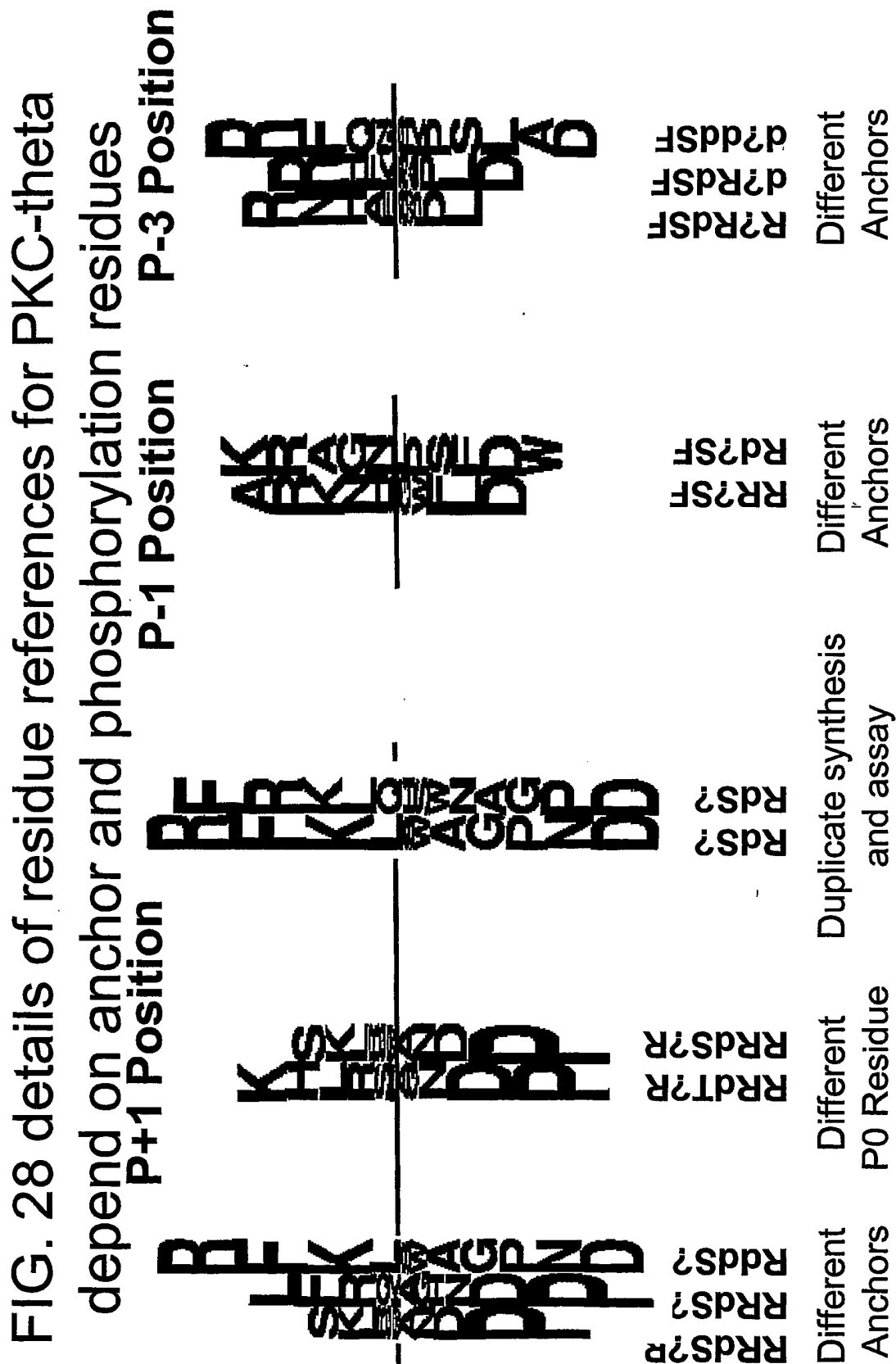constructs and altered conditions for kinase assay

FIG. 28 details of residue references for PKC-theta
depend on anchor and phosphorylation residues

FIG. 29 Results for ROK-alpha with test sets based on ??R??T???? with only 4 query residues
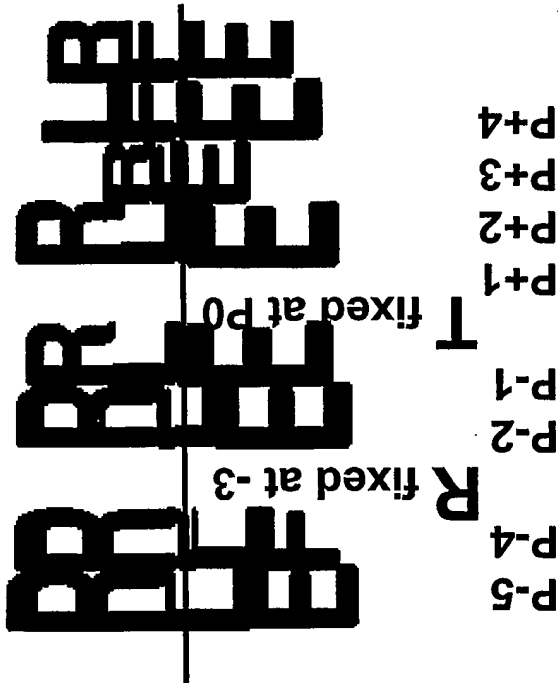
FIG. 30. Details of the R-Pair Anchor optimization set

Position of 1st R	Position of 2nd R				
	-6	-5	-4	-3	-2
-7	RRddddddSddd	RdRddddSddd	RddRddddSddd	RddddRdSddd	RddddRdsddd
-6		dRRddddSddd	dRddRddSddd	dRddRddSddd	dRddddRdSddd
-5			ddRRddddSddd	ddRddRddSddd	ddRddddRdSddd
-4				dddRRddSddd	dddRddRdSddd
-3					ddddRRdSddd

Position of 1st R	Position of 2nd R			
	-1	+1	+2	+3
-7	RddddddRSddd	RddddddSRdd	RddddddSdRd	RddddddSddR
-6	dRddddRSddd	dRddddSRdd	dRddddSdRd	dRddddSddR
-5	ddRddRSddd	ddRddSRdd	ddRddSdRd	ddRddSddR
-4	dddRdRSddd	dddRdSRdd	dddRdSdRd	dddRdSddR
-3	ddddRRSddd	ddddRSRdd	ddddRSdRd	ddddRSddR
-2	dddddRRSddd	dddddRSRdd	dddddRSdRd	dddddRSddR
-1		ddddddRSRdd	ddddddSRdd	ddddddSddR
+1			dddddddSRRd	dddddddSdRdR
+2				dddddddSdRR

FIG. 31: R-Pair set results for PKA

Position of 1st R	Position of 2nd R										
	-7	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**avg**
-7		1.0	0.3	-0.3	1.1	0.8	-0.8	-0.8	-0.4	-1.4	0.0
-6	1.0*		1.4	0.4	1.4	0.7	-0.5	0.1	-1.6	-0.7	0.2
-5	0.3	1.4		1.0	1.3	0.7	-0.4	-0.8	0.4	-0.9	0.3
-4	-0.3	0.4	1.0		1.9	0.8	-0.5	-0.5	-0.9	-1.0	0.1
-3	1.1	1.4	1.3	1.9		3.5	1.2	0.8	1.7	0.7	1.5
-2	0.8	0.7	0.7	0.8	3.5		1.3	-0.2	0.3	-0.1	0.9
-1	-0.8	-0.5	-0.4	-0.5	1.2	1.3		0.6	-1.4	-1.5	-0.2
+1	-0.8	0.1	-0.8	-0.5	0.8	-0.2	0.6		-0.2	-1.9	-0.3
+2	-0.4	-1.6	0.4	-0.9	1.7	0.3	-1.4	-0.2		-2.8	-0.5
+3	-1.4	-0.7	-0.9	-1.0	0.7	-0.1	-1.5	-1.9	-2.8		-1.1
avg	0.0	0.2	0.3	0.1	1.5	0.9	-0.2	-0.3	-0.5	-1.1	0.1

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

FIG. 32: R-Pair set reveals positions associated with the high preference for R
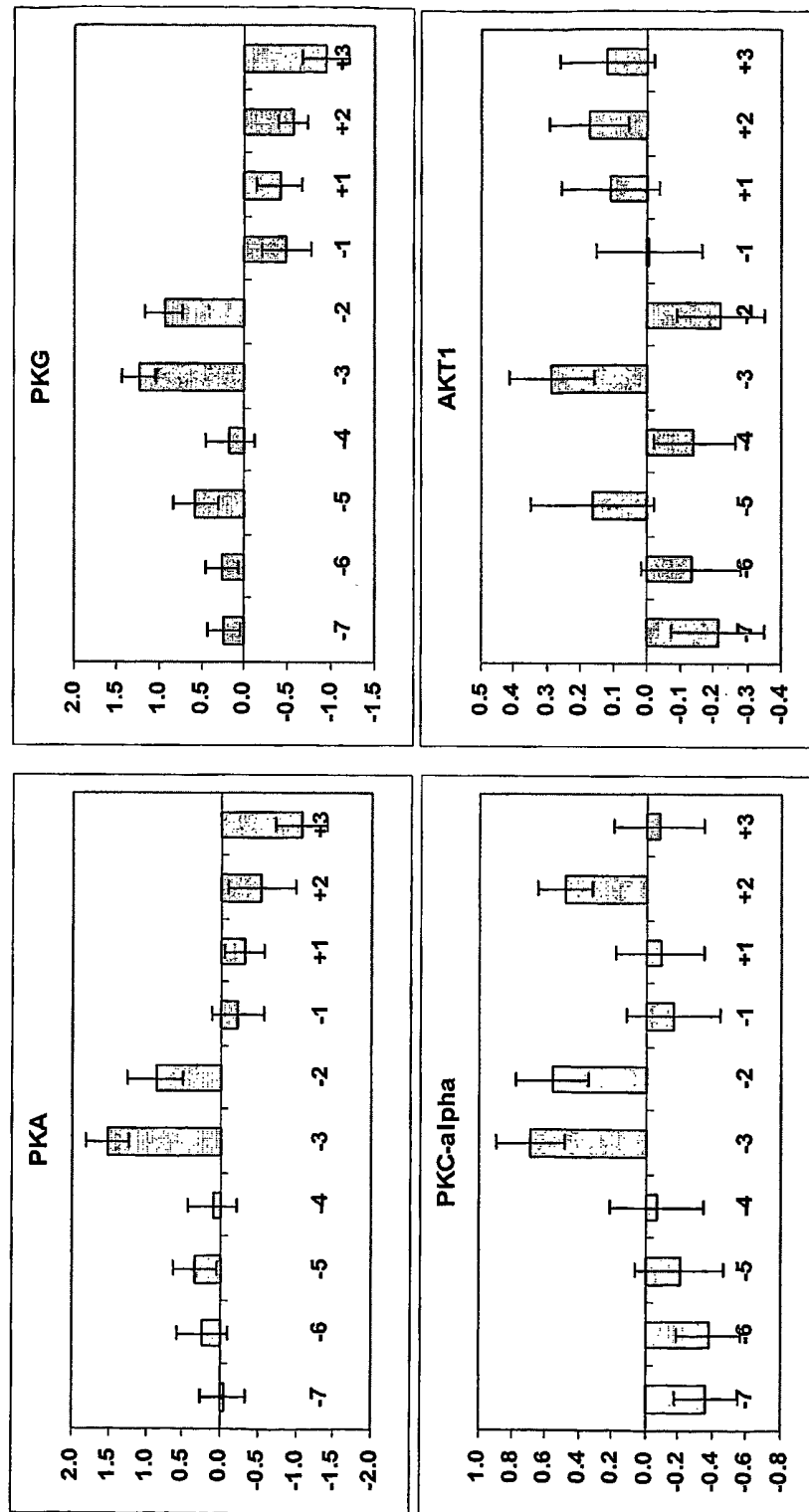
FIG. 33 : Detection of specific phosphorylation of LIMK-2 by Western blot with the pPKC antibody which is augmented following stimulation by the T-cell receptor
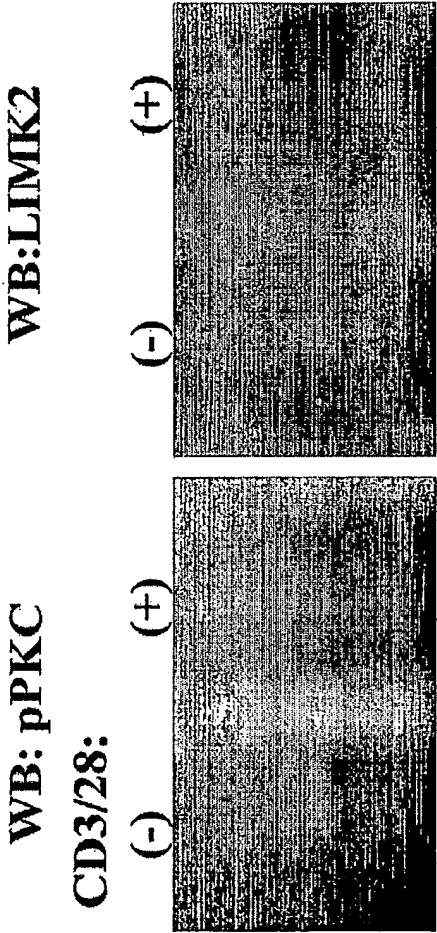

WB: pPKC

WB:LIMK2

CD3/28:

(-)   (+)

(-)   (+)

FIG. 34 : Detection of phosphorylation of MLK3 by Western blot with pPKC antibody
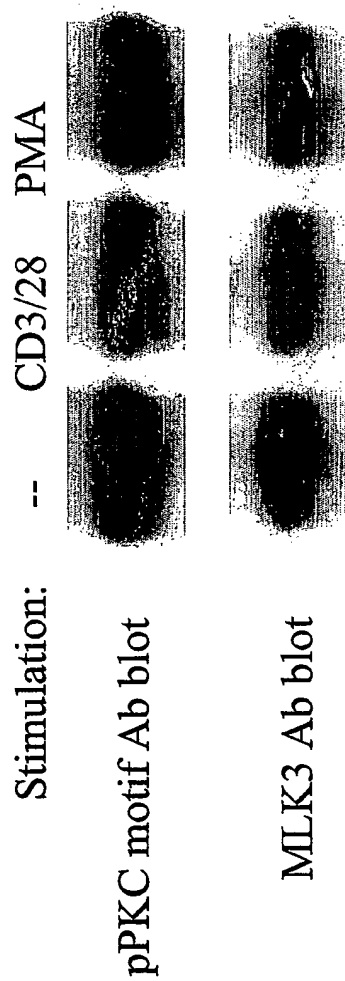
FIG. 35 Diagram of a computerized system in conjunction with which embodiments of the invention may be implemented
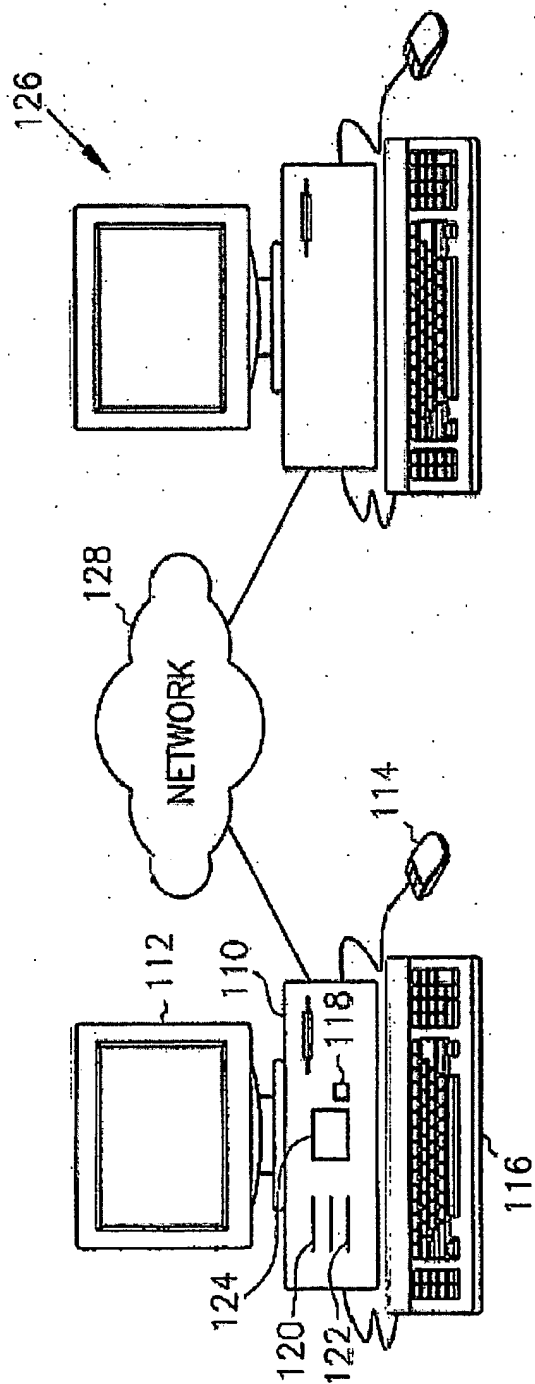
FIG. 36: RF-pair analysis for PKC-theta

-3R; -1F   BZGddddRddSFddd

-2R; +1F   BZGddddRdSFddd

F position	R position												F avg
	-7	-6	-5	-4	-3	-2	-1	P0	+1	+2	+3	+4	
-7		0.46	0.18	0.08	0.13	0.28	-0.36		-0.40	**0.54**	-0.06		0.09
-6	0.00*		0.22	-0.43	-0.09	**0.83**	**0.84**		0.20	-0.13	**-0.64**		0.09
-5	-0.66	**0.78**		-0.51	0.22	**1.68**	**0.71**		-0.45	0.01	0.06		0.21
-4	-1.11	0.35	-1.01		**0.75**	**1.86**	0.16		-0.39	0.31	-0.42		0.06
-3	-0.34	-0.57	-0.64	-0.36		0.36	-0.38		-0.08	0.02	-1.33	-0.73	-0.37
-2	-0.88	-0.43	-0.46	0.03	**0.60**		0.19		-0.58	-0.54	-0.69	-1.65	-0.31
-1	-0.79	-0.72	0.25	0.11	0.10	**0.65**			-0.55	-0.32	-0.80	-1.55	-0.23
P0													
+1	-0.18	**0.74**	**1.28**	**0.53**	**1.15**	**1.80**	**0.75**			**1.58**	**1.08**	0.20	**0.97**
+2	-0.13	0.40	-0.30	-0.16	**0.73**	**0.81**	-0.58		-0.95		-0.54	-0.76	-0.08
+3	0.46	-0.01	-0.37	0.37	-0.29	**0.69**	-0.63		0.14	0.24			0.07
R avg	0.11	0.11	-0.10	-0.04	0.37	1.00	0.08		-0.34	0.19	-0.37	-0.90	

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

FIG. 37: Average position-specific preferences of PKC-theta determined by RF-pair and R-pair set.
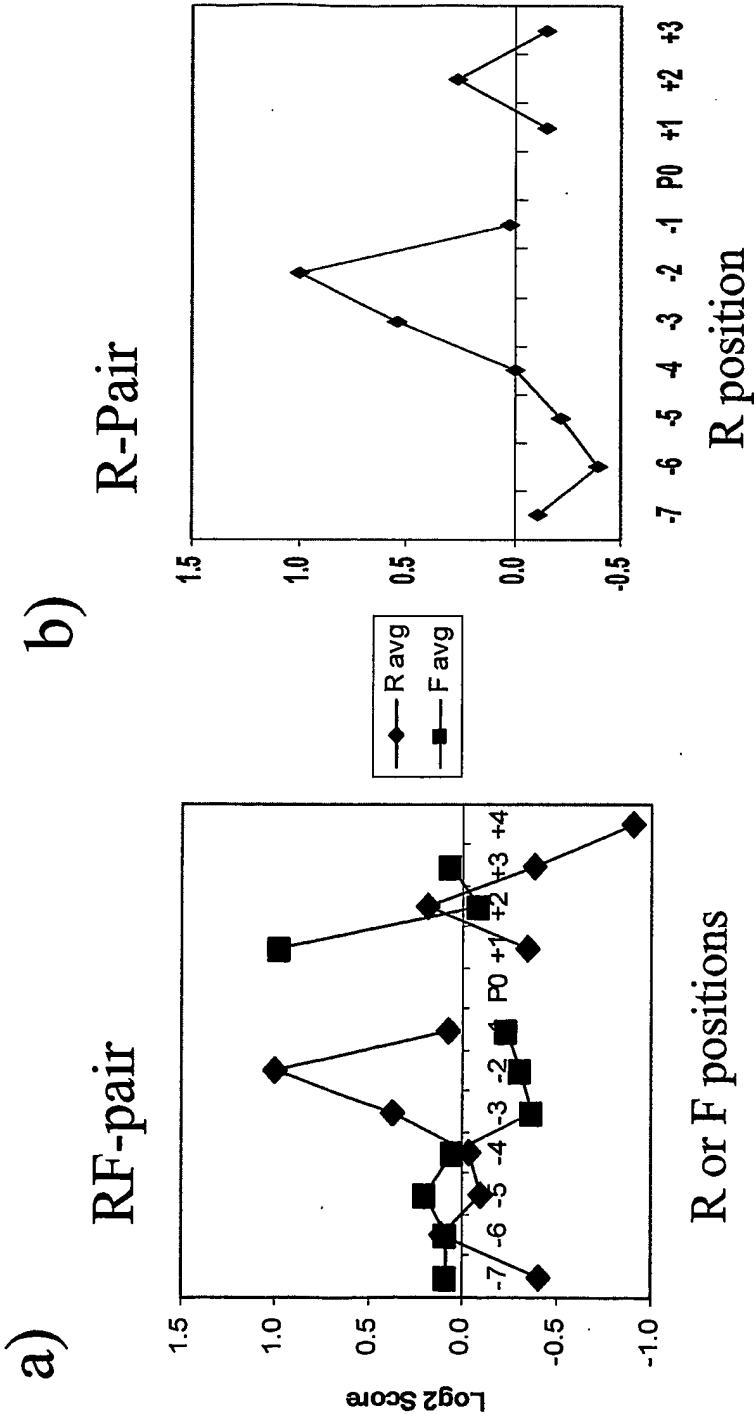
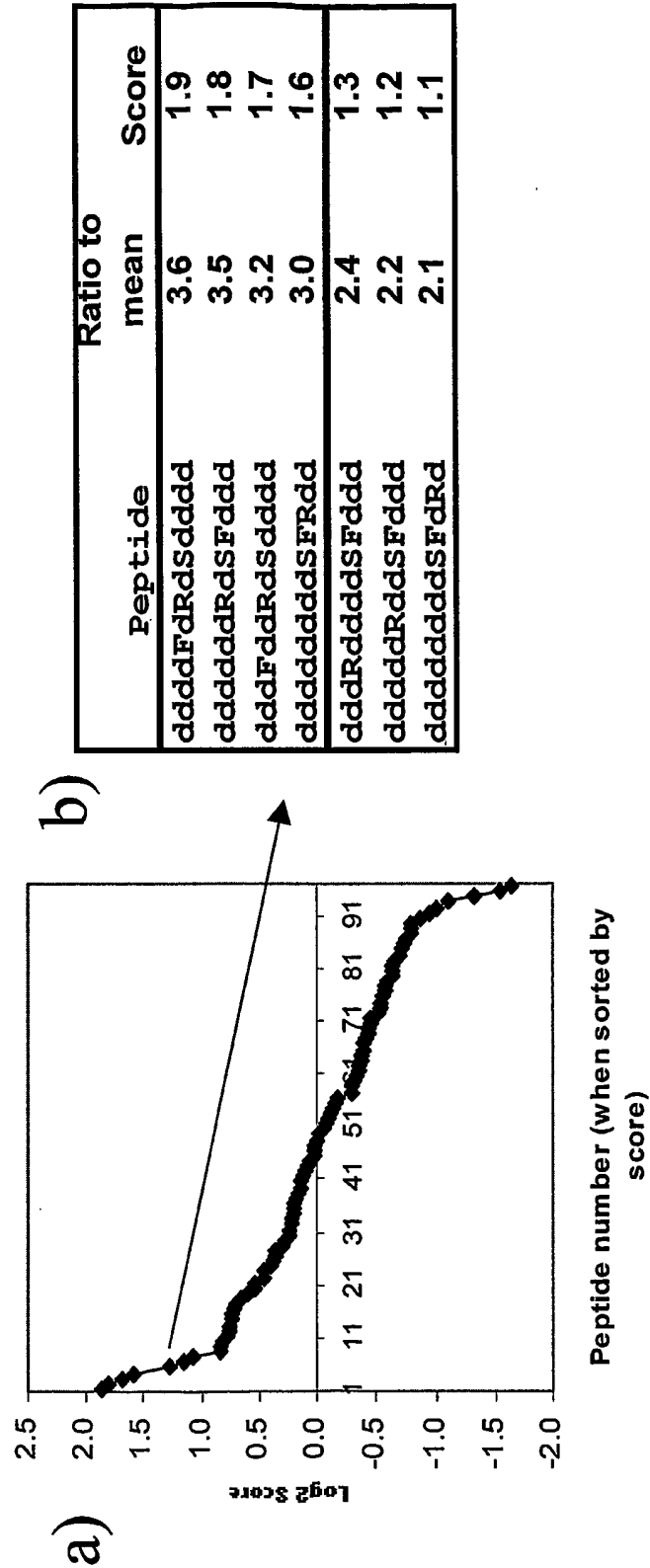# FIG. 38 More than one strongly preferred RF-pair peptide for PKC-theta

a)

b)

Peptide	Ratio to		Score
	mean		
dddFdRdSdddd	3.6		1.9
ddddddRdSFddd	3.5		1.8
dddFdddRdSdddd	3.2		1.7
dddddddSFRdd	3.0		1.6
dddRddddSFddd	2.4		1.3
ddddRdddSFddd	2.2		1.2
dddddddSFdRd	2.1		1.1

FIG. 39: PAK analysis with R-pair set

a)

		Position of 2nd R											
		-7	-6	-5	-4	-3	-2	-1	P0	+1	+2	+3	avg
Position of 1st R	-7		**0.6**	-0.1	0.4	-0.2	**1.1**	0.3		-0.4	-0.7	**-0.5**	0.0
	-6	**0.6***		0.0	0.4	0.5	**1.4**	0.2		**-0.7**	**-1.0**	**-0.8**	0.1
	-5	-0.1	0.0		**0.6**	**0.7**	**1.4**	0.2		-0.4	**-1.1**	**-0.9**	0.0
	-4	0.4	0.4	**0.6**		**0.7**	**1.7**	0.1		-0.3	-0.5	-0.1	0.3
	-3	-0.2	0.5	**0.7**	**0.7**		**2.4**	**0.9**		-0.2	-0.8	-0.3	0.4
	-2	**1.1**	**1.4**	**1.4**	**1.7**	**2.4**		**1.8**		0.0	-0.7	**0.7**	**1.1**
	-1	0.3	0.2	0.2	0.1	**0.9**	**1.8**			-0.6	-1.2	-0.4	0.1
	+1	-0.4	**-0.7**	-0.4	-0.3	-0.2	0.0	-0.6			0.2	**-0.9**	-0.4
	+2	**-0.7**	**-1.0**	**-1.1**	-0.5	**-0.8**	-0.7	-1.2		0.2		**-0.9**	**-0.7**
	+3	**-0.5**	**-0.8**	**-0.9**	-0.1	-0.3	**0.7**	-0.4		**-0.9**	**-0.9**		-0.4
	avg	0.0	0.1	0.0	0.3	0.4	**1.1**	0.1		-0.4	-0.7	-0.4	0.1

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.
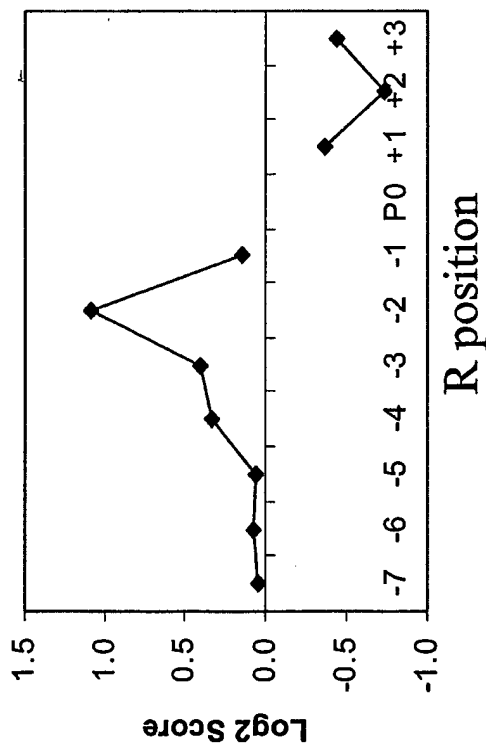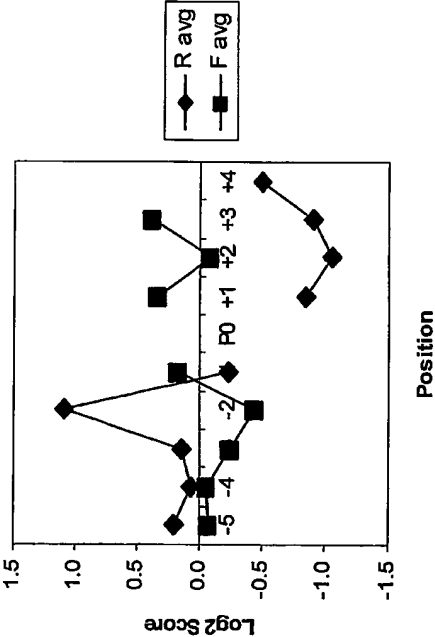
b)

Plot of Log2 Score versus R position (-7, -6, -5, -4, -3, -2, -1, P0, +1, +2, +3).

FIG. 40. PAK analysis with RF-pair

a)

F position	R position									
	-5	**-4**	**-3**	**-2**	**-1**	**P0**	**+1**	**+2**	**+3**	**+4**
-7	0.5	0.5	0.2	**1.2***	0.3	0.3	-0.1	-0.5	**-0.9**	**-0.8**
-6	0.3	0.3	0.2	**1.1**	0.0	0.0	**-1.0**	**-1.2**	**-1.1**	**-0.6**
-5		-0.2	-0.1	**1.3**	-0.3	-0.3	**-1.1**	**-1.3**	**-0.9**	
-4	0.1	0.1	-0.1	**1.0**	-0.3	-0.3	**-1.2**	**-1.1**	**-1.0**	
-3	0.0	0.0		0.4	-0.6	**-0.6**	**-1.4**	**-1.2**	**-1.4**	
-2	0.1	**-0.5**	**-0.5**		-0.8	**-0.8**	**-1.4**	**-1.4**	**-1.2**	
-1	0.5	0.1	0.4	**1.7**			**-0.9**	**-1.2**	**-0.8**	
P0										
+1	0.2	0.5	**0.6**	**1.1**	0.3	0.3	**-0.8**	**-1.2**	-0.3	-0.2
+2	-0.2	0.0	0.1	0.3	-0.4	-0.4			-0.5	-0.4
+3	0.4	0.5	**0.5**	**1.6**	-0.2	-0.2	0.3	-0.7		

*Log scores for substantially favored residues are bold underlined; log scores for substantially disfavored residues are bold without underline.

b)

Chart — Log2 Score (y-axis, 1.5 to -1.5) vs Position (x-axis: -5, -4, -2, P0, +1, +2, +3, +4)

Legend:
- R avg (◆)
- F avg (■)

FIG. 41. R position analysis for PAK using diverse basic proteomic set

a)

		R at P-3 Yes	R at P-3 No	R at P-2 Yes	R at P-2 No
phosphorylation > 10%	observed Yes	10	6	16	0
	observed No	45	35	38	42
	expected Yes	9.2	6.8	9.0	7.0
	expected No	45.8	34.2	45.0	35.0
	chi square	0.64		0.0001	

b)

Proposed site (aligned)	Relative Phosphorylation
WRKRM-S-LTLKS	100
SRNRL-S-SSS	69
ASVRK-S-FSED	39
RRQRQ-S-GVVVE	33
SLRRK-S-RSKD	31
PRGRR-S-RRAVS	30
RTRRF-S-KHLI	20
HRSRS-S-SNSSR	20
RERRR-S-YLTPV	17
DFLRR-S-SSRRN	17
RLSRR-S-RSASS	16
REKRR-S-TGVSF	16
YRHRG-S-LVNHR	14
REKRF-S-FFKKN	13
RKNRS-S-VKTPE	12
GGRRR-S-RSPDR	12

← P-2

c)

P-values for R at:								
P-6	P-5	P-4	P-3	P-2	P-1	P+1	P+2	P+3
0.08	0.04	0.01	0.64	0.0001	0.58	0.31	0.40	0.45

FIG. 42. pPKC antibody binding requires the SHP-1 residue S591 and constitutive active PKC-theta (PKC-theta CA) can promote its phosphorylation
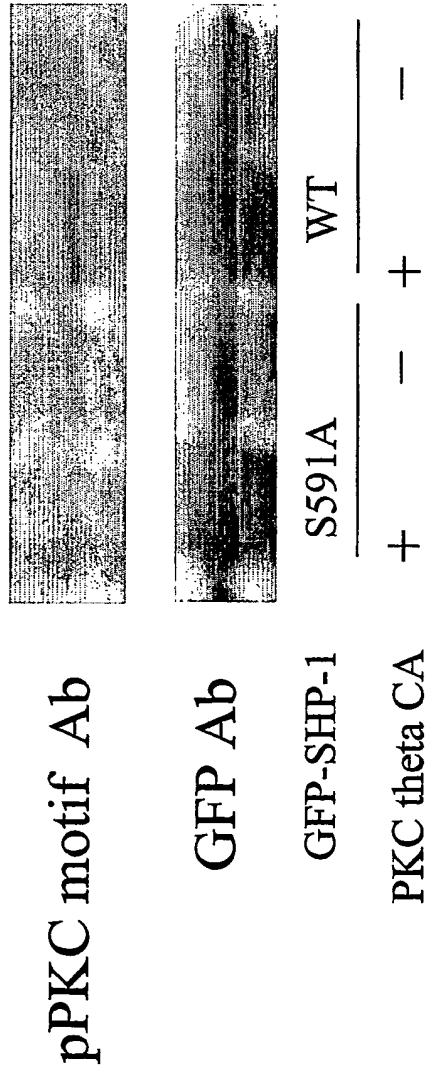
FIG. 43. SHP-1 S591 is phosphorylated in T-cells in response to CD3/28 or PMA
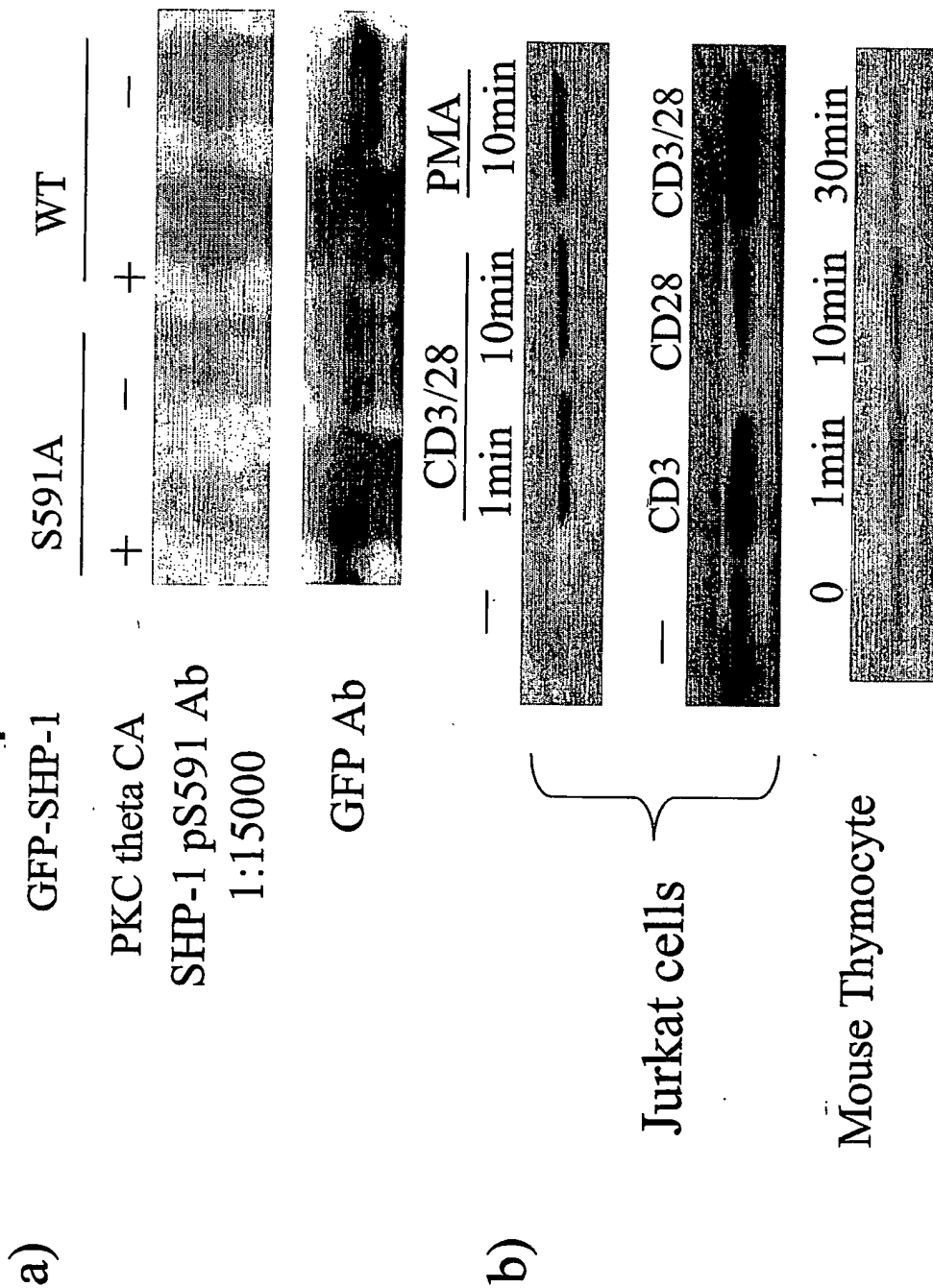
FIG. 44. PKC inhibitors interfere with phosphorylation of SHP-1 S591
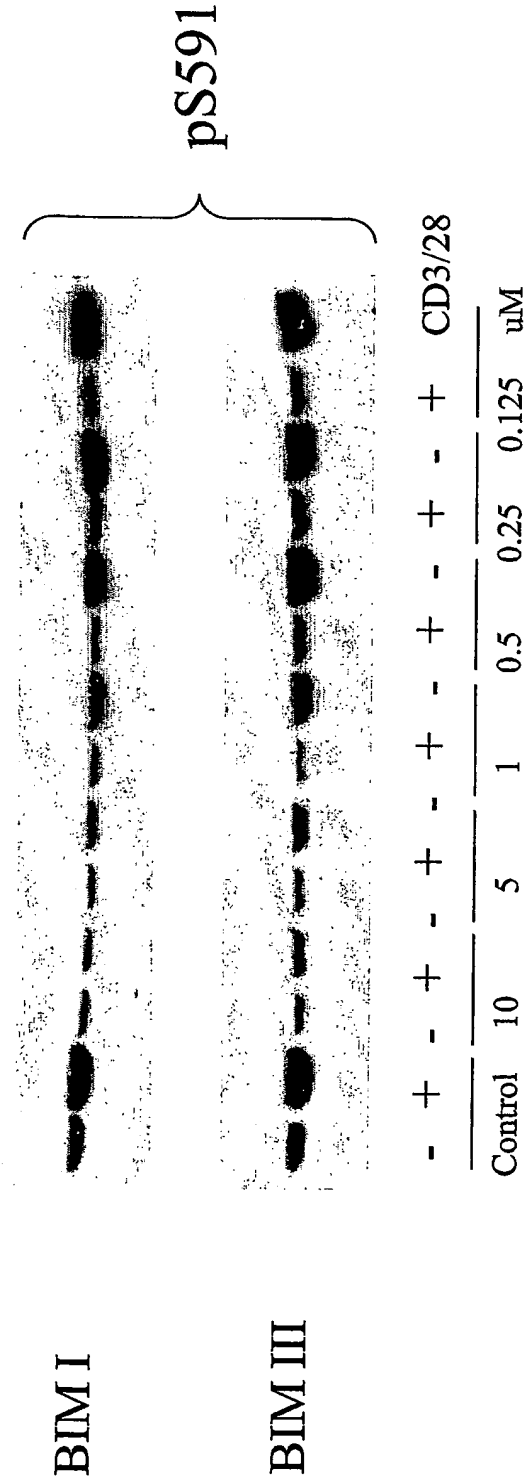
FIG. 45. Staining by anti-pS591 antibody is specific for SHP-1 Ser-591
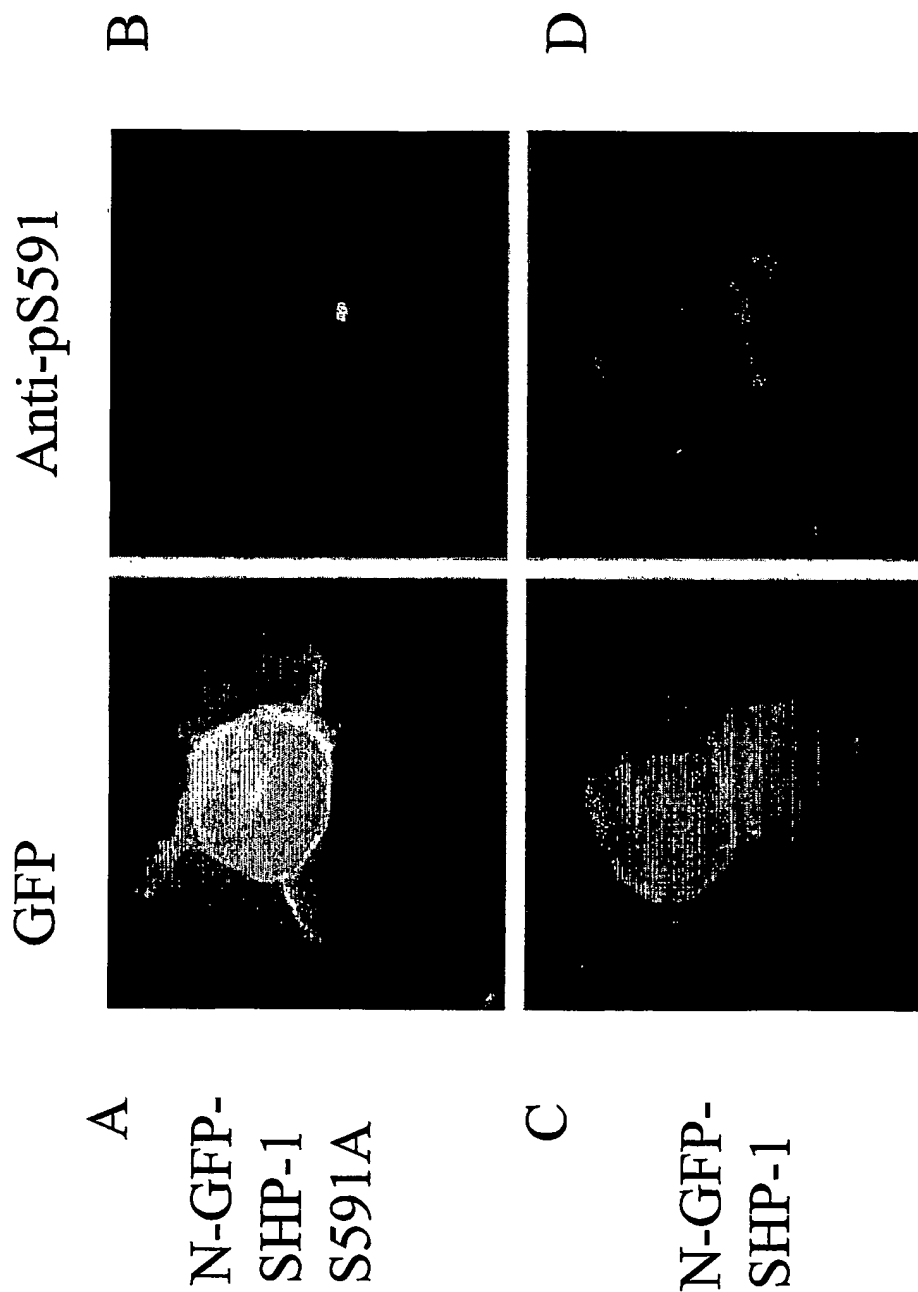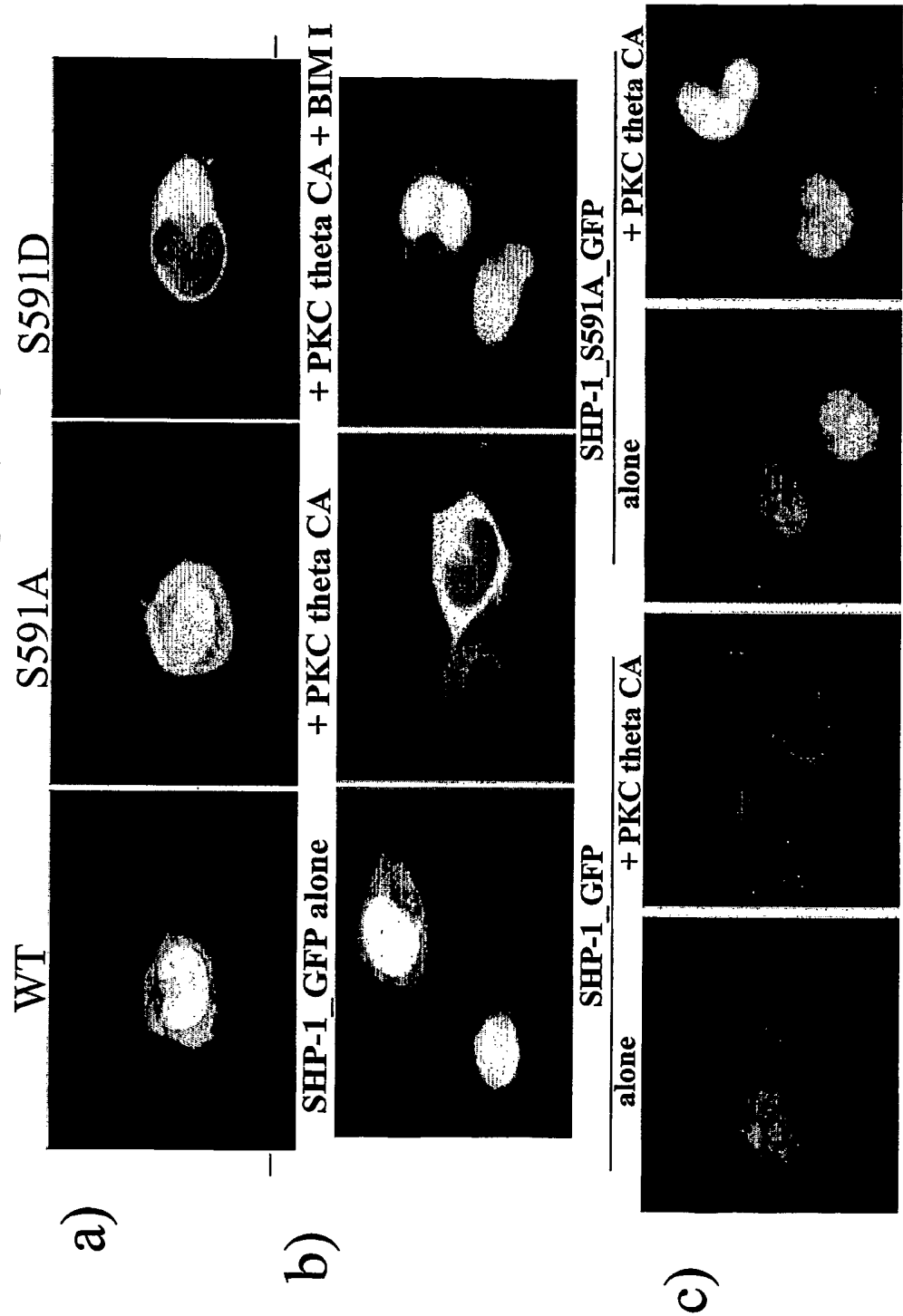
FIG. 46. Phosphorylation of S591 inhibits nuclear localization of SHP-1